A HYBRID ARCHITECTURE FOR RECOGNISING SPEECH SIGNALS IN MALAYALAM

Thesis submitted by

SONIA SUNNY

in partial fulfilment of the

requirements for the award of the degree of

DOCTOR OF PHILOSOPHY

Under the Faculty of Technology

DEPARTMENT OF COMPUTER SCIENCE

Cochin University of Science and Technology

Cochin - 682 022, Kerala, India

September 2013

A Hybrid Architecture for Recognising Speech Signals in Malayalam

Ph.D Thesis

Author:

Sonia Sunny
Department of Computer Science
Cochin University of Science and Technology
Cochin - 682 022, Kerala, India
sonia.deepak@yahoo.co.in

Supervisors:

Dr. David Peter S.
Professor in Computer Science & Engineering
Cochin University of Science and Technology
Cochin - 682 022, Kerala, India
davidpeter@cusat.ac.in

Dr. K. Poulose Jacob
Pro–Vice–Chancellor
Professor in Computer Science
Cochin University of Science and Technology
Cochin - 682 022, Kerala, India
kpj@cusat.ac.in

September 2013

Certificate

This is to certify that the work presented in this thesis entitled ***"A Hybrid Architecture for Recognising Speech Signals in Malayalam"*** *submitted to Cochin University of Science and Technology, in partial fulfilment of the requirements for the award of the Degree of* ***Doctor of Philosophy in Computer Science*** *is a bonafide record of research work done by* ***Ms. Sonia Sunny*** *in the Department of Computer Science, Cochin University of Science and Technology, under my supervision and guidance with Dr. K. Poulose Jacob, Pro-Vice-Chancellor, Cochin University of Science and Technology as Co-guide and the work has not been included in any other thesis submitted previously for the award of any degree.*

Kochi
September 2013

Dr. David Peter S.
(Supervising Guide)
Professor in Computer Science & Engineering
Cochin University of Science and Technology
Cochin – 682 022

Certificate

This is to certify that the work presented in this thesis entitled ***"A Hybrid Architecture for Recognising Speech Signals in Malayalam"*** *submitted to Cochin University of Science and Technology, in partial fulfilment of the requirements for the award of the Degree of* ***Doctor of Philosophy in Computer Science*** *is a bonafide record of research work done by* ***Ms. Sonia Sunny*** *in the Department of Computer Science, Cochin University of Science and Technology, under the supervision and guidance of Dr. David Peter S and myself, and the work has not been included in any other thesis submitted previously for the award of any degree.*

Kochi
September 2013

Dr. K. Poulose Jacob
(Co-Guide)
Pro-Vice-Chancellor
Professor in Computer Science
Cochin University of Science and Technology
Cochin – 682 022

Declaration

I hereby declare that the work presented in this thesis entitled ***"A Hybrid Architecture for Recognising Speech Signals in Malayalam"*** *submitted to Cochin University of Science and Technology, in partial fulfilment of the requirements for the award of the Degree of* ***Doctor of Philosophy in Computer Science*** *is a record of original and independent research work done by me under the supervision and guidance of Dr. David Peter S., Professor, Department of Computer Science and Engineering, Cochin University of Science and Technology and Dr. K. Poulose Jacob, Pro-Vice-Chancellor, Cochin University of Science and Technology. The results presented in this thesis have not been included in any other thesis submitted previously for the award of any degree.*

Kochi ***Sonia Sunny***

September 2013

Acknowledgments

The present study has been a process of invention and discovery, which has been a very challenging and reassuring academic exercise. This work would not have been what it is now, but for the unstinted support and guidance of many.

I thank God Almighty for the blessings showered on me, which has led to the completion of this research work.

I would like to express my sincere gratitude to Dr. David Peter S, Professor, Department of Computer Engineering, Cochin University of Science and Technology for his encouragement and guidance throughout the research work. His insightful comments, support and advice have been of immense benefit to me during the most difficult times.

I am grateful to Dr. K. Poulose Jacob, my co-guide and Pro-Vice-Chancellor of Cochin University of Science and Technology, for being a source of support and encouragement. His sincerity, patience and supportive attitude enabled the successful completion of this work.

My sincere thanks are also due to Dr. Sumam Mary Idicula, Head of the Department of Computer Science, Cochin university of Science and Technology for her support and guidance.

I would also like to thank Dr. K. Babu Joseph, Former Vice -Chancellor of Cochin University of Science and Technology, for his valuable help and advice.

I extend my deep gratitude to Rev. Dr. Harshajan Pazhayattil, Founder- Manager and Director, Prajyoti Niketan College for permitting me to avail the FIP Scheme and for pursuing this research project. His relentless pursuit of knowledge and zeal for perfection make me look up to him as a guiding light.

I wish to place on record my sincere thanks to Dr. Shaijan Paul, Principal, Prajyoti Niketan College, for his whole hearted support, motivation and understanding attitude.

I would like to thank Dr. Santhosh Kumar and Mr. Muralidharan - faculty members, Mr Joe Joseph - Librarian, technical staff members – Mr. Renjith, Mr. Shibu and Mrs. Manju, Mrs. Girija- Office staff and all the faculty members and office staff for providing the much needed help and support in completing the research work. My sincere thanks are also due to Dr. Sudheep Elayidom, Head, Department of Computer Engineering, Cochin University of Science & Technology for his timely advice and help.

I would like to extend my sincere thanks to my friends Ms. Anupama Surendran, Ms. Preetha Teresa Joy, Ms. Simili Joseph, Ms. Litta A.J and all the research scholars of my department for their cordiality, support and help. I am thankful to all my teachers at school and college, for making me what I am today.

I owe my deep-felt thanks to my colleagues at Prajyoti Niketan College-- Dr. Dhanya Menon for her help and support in editing this dissertation work and Dr. E Sandhya for her guidance in fine- tuning the various statistical methods. My sincere thanks to my colleagues of my department and all the teaching and non-teaching staff of Prajyoti Niketan College for their support. I would also like to place on record my gratitude to the students and staff of Cochin University, Prajyoti Niketan College and all who have contributed in developing the database and provided other related assistance.

Let me express my deepest gratitude to my husband Mr. Deepak George T, for his suggestions, help and moral support which brought this work to fruition. My sincere thanks to my children Diya – who helped me with much of the editing work-- Johan and Juan who looked after themselves during this difficult period, despite the lack of proper attention and care. I also extend my gratitude to my parents Prof. Sunny Thomas and Prof. Josamma Sunny and parents-in-law Mr. George Abraham T and Mrs. Ivy George for their encouragement, support and blessings. I also thank my relatives, friends and students for the concern and support they extended at various stages of my study.

Sonia Sunny

Abstract

Speech is the primary, most prominent and convenient means of communication in audible language. Through speech, people can express their thoughts, feelings or perceptions by the articulation of words. Human speech is a complex signal which is non stationary in nature. It consists of immensely rich information about the words spoken, accent, attitude of the speaker, expression, intention, sex, emotion as well as style. The main objective of Automatic Speech Recognition (ASR) is to identify whatever people speak by means of computer algorithms. This enables people to communicate with a computer in a natural spoken language.

Automatic recognition of speech by machines has been one of the most exciting, significant and challenging areas of research in the field of signal processing over the past five to six decades. Despite the developments and intensive research done in this area, the performance of ASR is still lower than that of speech recognition by humans and is yet to achieve a completely reliable performance level. The main objective of this thesis is to develop an efficient speech recognition system for recognising speaker independent isolated words in Malayalam.

Extensive work and research has been done in the field of speech recognition in different languages. But the performance of the speech recognition systems varies with factors like language, database used, number of speakers, differences among speakers, etc. So we are not in a position to categorically say whether one method is superior to another. Moreover, an ASR developed in a native language allows an illiterate person to interact with a computer. In this research work therefore, Malayalam, which belongs to the family of Dravidian languages of southern India and which also belongs to the family of classical languages with a rich set of alphabets is chosen for study. So, for this work, three new databases have been created for the Malayalam language since there are no built-in standard databases available in Malayalam. After creating the speech database, the speech recognition system proposed in this research work has been categorised under four independent modules. The different modules are:

a) Pre-processing of the speech signals
b) Extracting features from the signals
c) Post processing of the feature vector set obtained
d) Classification of speech features into appropriate classes

This research work is carried out in two phases. During the first phase of this research work, sixteen different experiments are conducted for developing speech recognition systems using different pre-processing, feature extraction, post processing and classification techniques. Different techniques like End Point Detection, Pre-emphasis Filtering, Frame Blocking, Windowing and Wavelet Denoising methods are deployed during the pre-processing stage. In the feature extraction stage, four techniques are used. Two spectral feature extraction methods namely Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficients (MFCC) as well as two wavelet based feature extraction techniques namely Discrete Wavelet Transforms (DWT) and Wavelet Packet Decomposition (WPD) are used for extracting the relevant features from the speech signals. In the post processing stage, the feature vectors are normalised using zero mean and unit variance. In order to analyse the performance of the speech recognition system, experiments are performed using four pattern classifiers namely Artificial Neural Networks (ANN), Support Vector Machines (SVM), Naive Bayes Classifiers and Hidden Markov Models (HMM). Then a comparison of the performance of these speech recognition systems in recognising the speech data stored in the Malayalam databases is carried out. From the results obtained, the speech recognition method which produced the optimum results for Malayalam language is elicited. The comparison of the performance of these developed systems is made on the basis of the degree of recognition accuracy.

Since the main objective of this work is to design a speech recognition system with utmost accuracy, new improved algorithms and modifications have been designed and developed for all the four modules of the speech recognition process during the second phase. From the results obtained from the comparison of these speech recognition systems developed, a new improved algorithm for feature extraction has been proposed and implemented by combining the two feature extraction techniques which produced the best recognition accuracies called Discrete

Wavelet Packet Decomposition (DWPD). A comparison of this hybrid method with the earlier methods was performed and the results obtained using the proposed methods have shown improvements in recognition accuracy. Speech signals are badly affected by background noise. This work proposes a new algorithm for smoothing a signal during the pre-processing stage before applying noise reduction techniques. Then a hybrid architecture is developed by combining this with wavelet denoising method based on soft thresholding called Adaptive Smoothing Soft Thresholding (ASST). A speech recognition system is developed using these smoothed signals and the DWPD method already developed. The results obtained using these are compared with the earlier results obtained in terms of Signal to Noise Ratio (SNR), Waveform Plots, Spectrograms, Recognition Accuracy, Precision and Recall. Results show that smoothing improves the recognition rate by improving the value of SNR.

This research work also proposes three statistical thresholding techniques for post processing the feature vector set obtained after feature extraction based on Three Sigma Limits, Quartiles and Confidence Interval Mode. The performances of these proposed methods are also found to be encouraging.

Finally, this work also proposes an ensemble of classifiers by combining the different classifiers using the techniques such as Bagging, Boosting and Stacking for the better performance of the speech recognition system. Further comparisons are made with the new ensemble classification systems developed and the systems that have already been developed. The results obtained show that each proposed algorithm developed during different stages of the speech recognition system yields better results. This in turn improves the overall performance of the system. Accuracy can be selectively varied depending on the applications and criticality of the tasks.

In recent years, various types of research in this specific area have been taking place in various languages. However, Malayalam speech recognition is still in its infant stage and only very few works have been reported in Malayalam. So developing an efficient speech recognition system for Malayalam is a challenging task. Bearing this in mind, the thesis proposes an improved speech recognition system for Malayalam.

Contents

List of Tables

List of Figures

Abbreviations

ASR	-	*Automatic Speech Recognition*
DSP	-	*Digital Signal Processing*
NLP	-	*Natural Language Processing*
LPC	-	*Linear Predictive Coding*
MFCC	-	*Mel Frequency Cepstral Coefficients*
DWT	-	*Discrete Wavelet Transforms*
WPD	-	*Wavelet Packet Decomposition*
ANN	-	*Artificial Neural Networks*
MLP	-	*Multilayer Perceptrons*
SVM	-	*Support Vector Machines*
HMM	-	*Hidden Markov Models*
PCA	-	*Principal Compound Analysis*
SNR	-	*Signal to Noise Ratio*
ZCR	-	*Zero Crossing Rate*
STE	-	*Short Time Energy*
NIST	-	*National Institute of Standards and Technology*
TIMIT	-	*Texas Instruments and the Massachusetts Institute of Technology*
SAAVB	-	*Saudi Accented Arabic Voice Bank*
RM	-	*Resource Management*
SUSAS	-	*Speech Under Simulated and Actual Stress*
IPA	-	*International Phonetic Alphabet*
IIR	-	*Infinite Impulse Response*
FIR	-	*Finite Impulse Response*
EPD	-	*End Point Detection*
PARCOR	-	*Partial Correlation Coefficients*
RBF	-	*Radial Basis Function*
FT	-	*Fourier Transforms*
FFT	-	*Fast Fourier Transform*

DCT	-	*Discrete Cosine Transform*
DFT	-	*Discrete Fourier Transforms*
MATLAB	-	*Matrix Laboratory*
WEKA	-	*Waikato Environment for Knowledge Analysis*
ARFF	-	*Attribute Relation File Format*
CSV	-	*Comma Separated Values*
STFT	-	*Short Time Fourier Transforms*
WFT	-	*Windowed Fourier Transform*
WT	-	*Wavelet Transforms*
MRA	-	*Multi Resolution Analysis*
db	-	*Daubechies Wavelets*
DWPD	-	*Discrete Wavelet Packet Decomposition*
ST	-	*Soft Thresholding*
HT	-	*Hard Thresholding*
AWGN	-	*Additive White Gaussian Noise*
Sym	-	*Symlet wavelets*
Coif	-	*Coifflets wavelets*
ASST	-	*Adaptive Smoothing Soft Thresholding*
DTW	-	*Dynamic Time Warping*
dB	-	*DeciBel*
K-NN	-	*K- Nearest Neighbour*
PNN	-	*Probabilistic Neural Network*
NN	-	*Neural Network*
ATM	-	*Automated Teller Machine*
TEO	-	*Teager Energy Operator*
PWP	-	*Perceptual Wavelet Packet*
SCWT	-	*Sampled Continuous Wavelet Transform*
PLP	-	*Perceptual Linear Prediction*
GMM	-	*Gaussian Mixture Model*

Chapter 1

INTRODUCTION

Contents

"All speech, written or spoken, is a dead language, until it finds a willing and prepared hearer."

Robert Louis Stevenson

1.1 Background

Speech processing and consequent recognition are important areas of Digital Signal Processing (DSP) as speech allows people to communicate more naturally and efficiently. Speech recognition is widely gaining attention because it allows a user to interact with the computer naturally without the use of a keyboard or any other interface. The main purpose of speech recognition is to extract the sequence of speech sounds and the message which best matches the input speech signal [1, 2]. ASR is basically a combination of various interdisciplinary technologies which include 1) Signal Processing, 2) Pattern Recognition, 3) Natural Language Processing (NLP) and 4) Linguistics which are unified into a statistical framework. Due to the developments in technology, better acoustic models, new feature extraction algorithms, better NLP techniques, DSP tools, as well as hardware and software tools, the

performance of ASR systems have steadily improved and there is much progress in the recognition of complex speech patterns. But this performance is still low when compared to the accuracy and speed with which human beings recognise speech especially when there exist adverse situations like background noise, disturbances and other degradable conditions that affect the quality of speech signals [3]. This has caused a considerable performance gap between humans and machines [4, 5]. So more research and developments are needed in this area so as to enable a computer to achieve the speech recognition ability similar to that of human beings.

We live in an era in which a lot of importance is being accorded to individual recognition systems based on physical parameters. Speech with its inherent variations and diversity is one of the foremost and conclusive techniques in recognising an individual through the process of verbal signs. There has been a lot of research in the area of speech recognition for different languages like English, Chinese, Arabic, Turkish, Bengali, Hindi, Tamil, etc. But only very few works have been reported in Malayalam. Malayalam, which belongs to one of the four major Dravidian languages of southern India and is the native language of Kerala, is chosen for our study. So developing an efficient speech recognition system which has more ability to recognise speech is of great importance and is a significant and challenging area of research.

1.2 Motivation

Now-a-days, speech recognition is extensively used due to its versatile and wide range of applications. Speech recognition is applied in almost all fields of life such as mobile applications, weather forecasting, agriculture, healthcare, military, database querying, automatic voice translation, command and control, training air traffic controllers, telephone directory assistance,

office dictation devices, robotics, video games, transcription, etc [6]. Moreover, transferring data to a computer through spoken language is much faster than that of transferring data through hand writing or using keyboard.

Today, most of the technologies related to human machine interaction are limited to people who can read, write and have a minimum knowledge especially of English. So, spoken language interfaces to computers are a topic that has lured and fascinated engineers and speech scientists. But this also needs some knowledge about the language. However, if it is possible to interact with a machine in a native spoken language, then it enables more people, especially illiterate people, to benefit from the advantages of Information Technology. Now-a-days, people have realised the need for an efficient human machine interaction system based on Indian languages. So, researchers are currently striving hard to improve the accuracy of the speech processing techniques in these languages so as to facilitate a user friendly interactive system [2]. If an efficient speech recogniser is developed, a natural human-machine interface can be achieved. A person can naturally and easily use the computer without any special tools or devices. Moreover, such a system can be used by any person who is able to speak and this will allow an even broader use of machines, specifically computers. Researchers are giving more attention in the area of speech recognition particularly because voice recognition may develop as the primary user interface in future.

Of the various native languages available, Malayalam is a language spoken in India and is the official language of the state of Kerala. It belongs to one of the 22 scheduled languages of India. It is spoken by about 38 million people mainly in the state of Kerala and the union territories of Lakshadweep and Puducherry. Malayalam is a language with a number of spoken Sanskrit words. Malayalam is the youngest of all developed languages of the Dravidian

family. With the Cabinet's decision, Malayalam is now a classical language and joins Tamil, Kannada and Telugu, all members of the Dravidian linguistic family, as the fourth 'classical' language in South India. A classical language is a language with a literature that is classical. According to UC Berkeley linguist George L. Hart, "*it should be ancient, it should be an independent tradition that arose mostly on its own, not as an offshoot of another tradition, and it must have a large and extremely rich body of ancient literature*" [7].

The Malayalam language has a rich set of alphabets which consists of 51 alphabets out of which 15 are vowels and 36 are consonants. Malayalam is one of the richest languages in terms of number of alphabets. Moreover, it is one of the toughest languages when it comes to speech considering the variations in its alphabets which are very subtle in pronunciation. A working model for the language will be a comprehensive work which then can be easily replicated to other languages. A system which can engage in a dialogue with the users in native language provides empowerment to illiterate people living in rural area. Farmers can access information about daily prices for crops, agricultural techniques, and daily local news bulletins even through communication using spoken words. So, developing an efficient speech recognition system in Malayalam has great relevance.

Since speaking is a more convenient and easy way to communicate with a computer than typing, designing a speech recognition system which can perform accurately is of immense importance. While designing a speech recognition system, several things are to be taken into consideration like availability of a good database, defining the speech classes, selection of signal pre-processing methods, feature extraction techniques adopted, post processing methods used, speech classifiers used and the performance evaluation methods adopted [8]. Speech recognition systems being a

combination of these different modules, improvements in all these stages are needed to build an efficient system. So it is necessary to design a system which can perform better than the existing systems by designing new algorithms and models which are capable of improving the performance of the different modules used. Thus the main motivation behind this research work is to improve the performance of different stages of a speech recognition system, thereby improving the overall performance of the speech recognition system developed in Malayalam.

1.3 Problem Statement

The main objective of ASR is to develop an efficient and a more accurate mechanism to recognise human speech using different algorithms executed in a computer. ASR is essentially a sequential pattern recognition problem which includes multi-classes. The relevant features from the speech signals which form the feature vector set are extracted and these are classified as the corresponding speech samples [9]. There are several challenges involved in developing an ASR and hence recognising speech is a very complex task. Speech recognition is a difficult task since there are many factors which limit the performance of a speech recognition system [10]. The main limitations are variability in channel, speaker, speaking style, spoken language, sex of the speaker, regional and social dialects, speed of speech, ambiguity in speech, phonetic identity, pitch, microphone and presence of background noise.

In spite of these limitations and variability, the main research question posed in this work is this: "*How can a speech recognition system, for recognising speaker independent isolated words in Malayalam with maximum recognition accuracy, be designed?"* This work aims at improving the

performance of a speech recognition system by developing new algorithms and by introducing improved methods in the different stages of a speech recognition process. In order to tackle this challenging research problem, the following sub questions are to be considered.

i. Which is the best feature extraction method and classifier combination that is appropriate for the recognition of words in Malayalam?

ii. What are the steps to be taken to improve the performance of the speech recognition models?

iii. How can the noise in a speech signal be reduced?

iv. How can a better feature extraction technique be designed?

v. How can a compact and consistent set of features for better classification be selected?

vi. How can the performance of the classifiers be improved?

1.4 Research Goal and Objectives

The ultimate goal of this research work is to design an efficient speech recognition system which has improved performance based on recognition accuracy for recognising speech. Speech recognition involves a fusion of several information sources. It can also be considered as a pattern recognition problem since the speech samples are classified into different classes according to the words spoken. This research work focuses on the exploitation of various state-of-the-art ASR techniques for recognising speech in Malayalam. It also compares the performance of these techniques and designs new enhanced techniques for the better performance of the speech recognition system. In order to achieve this goal, the objectives of the work are identified as follows:

i. Identify the combination of feature extraction method and classifier which produces the optimal results for the databases created in Malayalam.
ii. Develop a new improved feature extraction method for improving the optimal results obtained.
iii. Create and implement a new enhanced algorithm for removing noise from the speech signal.
iv. Design techniques for generating a more consistent feature vector set for better classification results.
v. Adopt methods for enhancing the performance of pattern classifiers.
vi. Evaluate the performance of each new algorithm/method developed during each stage of the speech recognition process.

1.5 Contribution of the Research Work

The main contributions of this work can be summarised as follows.

1. Three databases for vowels, digits and isolated words are created for the Malayalam language.

2. Sixteen different speech recognition systems are designed, developed, implemented and tested for these datasets. The speech recognition systems are developed with different phases which include signal pre-processing, feature extraction, post processing of the feature vectors, training and classification or recognition.

3. The performance of four major feature extraction techniques namely LPC, MFCC, DWT and WPD in extracting the relevant features from the speech samples is evaluated.

4. The performance analysis of four pattern classifiers namely ANN, SVM, HMM and Naive Bayes Classifier in recognising the speech signals is carried out.
5. An optimal speech recognition system is chosen for Malayalam by comparing the performance of the combination of the above feature extraction techniques and the classifiers.
6. A new improved hybrid algorithm is developed for feature extraction which improves the performance in terms of recognition accuracy.
7. A new algorithm is developed to smooth the signals so that the sudden spikes present in the speech signals are reduced thereby reducing the noise contents which in turn improves the recognition accuracy.
8. Three statistical thresholding techniques are implemented in the feature vector set obtained to select and transform the feature vectors to a format which is more suitable for further classification based on:
    a) Three Sigma Limits
    b) Quartiles
    c) Confidence Interval Mode
9. Ensemble of classifiers are carried out by combining different classifiers to improve the overall performance of the speech recognition system using three techniques namely:
    a) Bagging
    b) Boosting
    c) Stacking.

1.6 Outline of the Thesis

The complete approach of the research reported in this thesis is described in figure 1.1 given below which consists of 9 chapters followed by a brief description of the contents in each chapter.

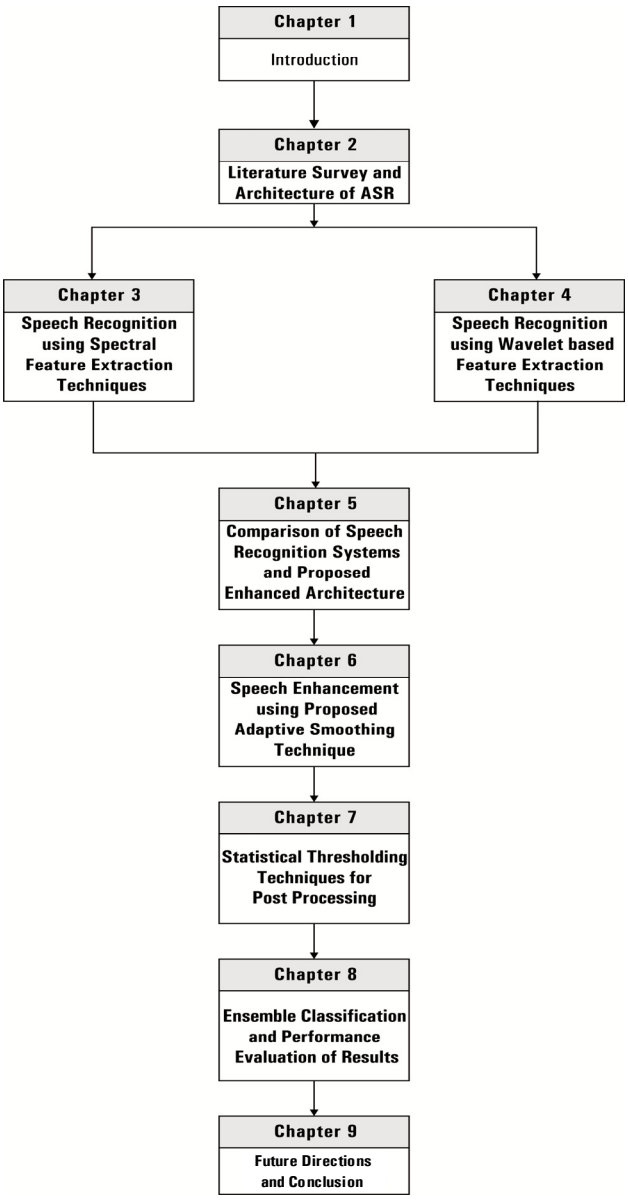


Figure 1.1: Structure of the Thesis

Chapter 1: This chapter provides an introduction to this research work which includes the background, motivation, objectives and main contributions of this work.

Chapter 2: This provides a brief survey of a selection of previous work done in the area of speech recognition in various languages, different techniques available for the recognition of speech signals and their performance. This also includes the architecture of the speech recognition system developed and the different stages in the speech recognition process.

Chapter 3: This chapter focuses on the spectral feature extraction methods LPC and MFCC. This includes a brief review of the steps in the feature extraction process and its implementation in the Malayalam databases created.

Chapter 4: This chapter demonstrates the speech recognition process using wavelet based feature extraction techniques namely DWT and WPD. It also gives a description about the various wavelet families and the implementation of these techniques in the databases created.

Chapter 5: This includes a summary of the performance evaluation of the above mentioned methods to identify the best feature extraction and classifier combination with optimal recognition rate. This chapter also proposes a new improved feature extraction method for improving the performance of the speech recognition system.

Chapter 6: This chapter presents a novel method for enhancing speech signals. Here a new algorithm is proposed and implemented for smoothing the signals to remove sudden spikes in the signal thereby reducing noise.

Chapter 7: This deals with the application of three statistical thresholding techniques namely Three Sigma Limits, Quartiles and Confidence Interval

Mode in the feature vector set during the post processing stage to select the best features to improve the recognition accuracy.

Chapter 8: This chapter presents the ensemble classification approach which combines the different classifier models to improve recognition accuracy. A comparative study of the performance of these ensemble classification techniques namely Bagging, Boosting, Stacking and all the speech recognition systems developed using different methods is also evaluated.

Chapter 9: This chapter concludes the thesis by presenting the highlighting features of the work. It also discusses the future directions for extending the research work.

References are listed after Chapter 9 along with the details of **publications** made by the author.

1.7 Summary of the Chapter

The background, motivation, scope and objectives of this research work are explained in this chapter. The chapter concludes by pointing out the main contributions of this research work. An outline of the succeeding chapters is also presented.

Chapter 2

LITERATURE SURVEY AND ARCHITECTURE OF ASR

Contents

As discussed in the previous chapter, the main objective of this work is to design an efficient speech recognition system for Malayalam speech. So this chapter presents the main framework of the speech recognition process. An overview of the speech production mechanism along with the different approaches to speech recognition process is explained here. This chapter also presents a survey of the literature on speech recognition systems that includes the various research works done in different languages. A brief description of the architecture of the speech recognition system including the description of the databases created in Malayalam is explained in this chapter.

2.1 Introduction

The main function and intention of speech is communication and it is still the first and foremost means of communication. A speech signal is a one dimensional stream of data. Unlike other signals which are stationary in nature, speech is non stationary in nature where the frequency changes with time. So, different approaches and methods are essential for the processing of

a speech signal. Since speech signals are time-varying in nature, a time-frequency analysis of the signal is necessary.

Speech signals demonstrate a range of inter-speaker variations for the same utterance in the time domain. A speech signal is characterised by a number of features like physical and perceptive features. Physical features represent the physical characteristics of a speech signal like energy, Zero Crossing Rate (ZCR) and its related features, fundamental frequency, spectral features like bandwidth, formant location as well as time domain features like duration of the sample. Perceptive features are based on how human beings perceive the sound signals. The important perceptive features are pitch and prosody, rhythm, timbre, voiced/unvoiced frames etc [11]. So extracting the most relevant features plays an important role in determining the performance of a speech recognition system. By using an appropriate classifier for these features, a speech recognition system can be developed. The choice of a suitable combination of feature extraction method and classifier, therefore, has a significant role in determining the efficiency of the speech recognition system.

Section 2.2 provides a brief overview of the speech production process. The different approaches to ASR are discussed in section 2.3. Section 2.4 presents a survey on various speech recognition systems developed in different languages. Architecture of the speech recognition system developed for finding the best features and classifier combination and the different stages in the development process are explained in section 2.5. The chapter is concluded in section 2.6.

2.2 Speech Production

Speech sounds are produced when air flows from the lungs and passes the vocal tract and then through the throat and mouth. The lungs are

considered as the source of the sound and the vocal tract is considered as the filter that produces different types of sounds [12]. From a simple engineering point of view, speech production can be considered as an acoustic filtering process in which a speech sound source excites the vocal tract filter. Many organs are involved in the production of speech - lungs, larynx, vocal cords, uvula, palate, tongue, teeth, lips, nose and different parts of the mouth [1]. Since speech is produced as a sequence of sounds, the state of the vocal cords and the shape, size and position of the various articulators change over time depending on the sound being produced [2].

There are many factors that affect the production of sound. The different sounds produced depend on the vibration/lack of vibration of the vocal cords, shape of the vocal tract, the amount of air that is pulled into the lungs and the nature of the source - whether periodic, noisy, impulsive, or a combination of these three [13, 14]. The speech signals are classified according on whether a sound is produced, or if produced, whether the vocal cords are vibrating or not [12, 15]. According to these criteria, speech can be classified as:

a) **Voiced:** Vocal cords are tensed and vibrate periodically resulting in a quasi-periodic speech waveform.

b) **Unvoiced:** Vocal cords do not vibrate. So there is no periodic random speech waveform.

c) **Silence**: No speech is produced.

2.3 Approaches to Automatic Speech Recognition

There have been a number of advances in the field of speech recognition in the past few decades which, have led to tremendous improvements in the implementation of a comprehensive speech recognition

system. Almost all the works done in ASR are based on the three main approaches to speech recognition namely *a) Acoustic Phonetic approach b) Pattern Recognition approach and c) Artificial Intelligence approach.*

2.3.1 Acoustic Phonetic Approach

In this approach, it is assumed that a spoken language is made up of finite, distinctive phonetic units with a set of acoustic properties. There are three steps in this approach [12, 16].

- *Spectral analysis of the speech signal* which converts the spectral measurements into a set of features.
- *Segmentation and labelling* where the speech signal is divided into different segments and phonetic labels are attached to these segments.
- *Determination of a valid word* from the phonetic labels generated during the second step.

The most common spectral analysis methods are *Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficients (MFCC)* methods.

2.3.2 Pattern Recognition Approach

This is a commonly used technique which has two steps [12, 17, 18].

- *Pattern training* where training algorithms are used to train the patterns.
- *Pattern comparison* where the speech to be recognised is compared with patterns which are already trained and learned in the training stage.

2.3.3 Artificial Intelligence Approach

This is a hybrid approach of combining acoustic phonetic and pattern recognition approaches [12, 19]. It exploits the concepts of these two methods. This method attempts to mechanise the recognition procedure in a manner which

is similar to how a person applies his intelligence in visualising, analysing and taking decisions on the measured features. In this approach, knowledge is incorporated from different knowledge sources and is applied on the problem [12].

2.4 Literature Survey on Speech Recognition Systems

Now-a-days, it is evident that ASR has great importance in almost all fields due to its wide range of applications. So, a number of researchers are involved in this challenging problem of developing speech recognition systems with utmost accuracy. Since speech is the most natural and easiest way of communication, speech recognition has become an exciting area of research. So designing a machine which can mimic human behaviour by responding to spoken language is a very significant field of study. For this, different combinations of acoustic, articulatory, and auditory features are used in many of the speech recognition systems [1]. In recent years ASR has reached very high levels of performance, by dropping the error rates and thereby increasing the recognition rate. This current state of performance is largely due to improvements in different areas like the availability of common speech corpora which allows the easy use of large training sets, new ideas in acoustic modeling, the use of statistical techniques and improvements in search algorithms. Despite this, there is vast scope for improvement in developing a reliable system.

A speech recognition process involves the selection of an appropriate set of features and the classification of these feature vectors into corresponding speech classes. Since there exist different types of feature extraction techniques [20] and pattern classifiers which play an important role in obtaining good recognition rate, development of these are to be reviewed to

find out the best techniques which are more appropriate for a speech recognition system [21]. If the features extracted from speech samples are poor, then even the best classifier cannot recognise the speech samples correctly. Likewise, if the features extracted are relevant, but a poor classifier is used for pattern classification, then it will generate poor results. So, it is necessary to have knowledge about the merits and demerits of both feature extraction methods and pattern classifiers which can be attained only through a proper review of the previous works done in this field.

The idea of speech recognition by a machine came into existence in the early 1920s. Since then, there have been remarkable progress in the field of speech recognition which have led to tremendous improvements in the implementation of a comprehensive speech recognition system. A brief history of the technological advances in the field of ASR from 1920s is given in [15, 22, 23, 24, 25, 26, 27, 28]. Though research in the field of speech recognition started in the 1920s, major advances were made from the 1960s onwards with advanced speech representations like spectral analysis methods, statistical methods and wavelet- based techniques.

During the last five decades, different categories of speech recognition systems have come into existence. Speech recognition systems can be classified into various types depending on *the type of speech utterance, type of speaker model and the type of vocabulary* [21]. Depending on the *type of speech utterance*, speech can be classified into:

a) Isolated words **–** where there is a pause before and after each word and the system is designed to identify single words at a time.

b) Discontinuous speech **-** which consists of full sentences in which words are artificially separated by silence.

c) Continuous speech - which are naturally spoken sentences and

d) Spontaneous speech - which includes natural unrehearsed speech.

Based on *speaker models*, speech- recognition systems can be divided into:

a) Speaker - independent models - that recognise the speech patterns of a large group of people.

b) Speaker - dependent models - which are designed to recognise speech patterns from a single person and

c) Speaker - adaptive models - which usually begins with a speaker independent model and adjusts these models more closely to each individual during a brief training period.

The *size of vocabulary* of a speech recognition system also affects the complexity, processing requirements and the accuracy of the system [8]. According to this criterion, the different classifications are:

a) Small Vocabulary Systems - that require only few words.

b) Large Vocabulary Systems - that need large number of vocabularies.

Speech recognition therefore, ranges from recognition of isolated words to continuous speech recognition, speaker-dependent systems to speaker - independent systems and small vocabulary to large vocabulary systems. In this research work, a speech recognition system is developed for ***speaker independent isolated words with medium vocabulary.*** Since a speaker - independent speech recognition system is developed, the speaker specific characteristics of the speech signal are ignored [2]. Moreover since there exists multitude of languages the world over, a literature survey on various types of research carried out in each of them is a virtually impossible task. Each language has its own uniqueness making each research distinct.

A standard database is the primary requirement and its quality is an integral factor for the proper recognition of speech samples. Standard databases for the following are known to exist in many of the languages in the following linguistic contexts:

a) *for speaker dependent and speaker independent speech data*
b) *isolated words and spoken sentence databases*
c) *digits databases*
d) *databases with and without noise contents*
e) *databases for speaker identification*
f) *databases for identifying the language*
g) *native and non-native speech databases*
h) *databases with telephone conversations and*
i) *small and large vocabulary databases*

One of the most popular speech database is the TIMIT database which contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. Other TIMIT related corpora include CTIMIT, FEMTIMIT, HTIMIT, NTIMIT, etc. which were recorded using different recording input devices such as telephone handset lines and cellular telephone handset. VidTIMIT is a database which consists of both video and corresponding audio recordings. TIDigits is a database for digits and Switchboard is a large multi speaker corpus of telephone conversations. Some of the other standard databases are NIST for speaker recognition, RM1 and RM2 for continuous speech recognition with speaker independent and speaker dependent data set, SIVA, Polycost and YOHO for speaker recognition, etc [8]. There exists a variety of non native speech databases of non-native pronunciations of English like Jupiter with

telephone speech, IBM-Fischer for digits, PF- STAR with children's speech, etc., SUSAS database with speech containing stress and emotions, the telephony Arabic speech corpus for isolated digits named SAAVB, Hindi Speech corpus for Hindi, Chinese Mandarin Speech Recognition Database, Canadian French Speech Recognition Database, Russian Speech Recognition Database, Japanese Speech Recognition Database, Polish Speech Recognition Database, US English Speech Recognition Database, etc. and researches on standard databases are found to perform better.

This section provides a brief survey of some of the recent works done in recognising *speaker independent words, digits and vowels in different languages.* Table 2.1 given below gives a summary of the previous works done in multiple languages.

Table 2.1 List of previous works done based on sample space and nature of the database, features, classifier, language and accuracy.

Author	Sample Space & Nature of the database	Features	Classifier	Language	Accuracy
Md Salam et al. [29]	4 speakers, 10 digits, 20 utterances per speaker	LPC	MLP	Malay	95%
Thiang et al. [30]	30 speakers, 7 words for robot, single utterance per speaker	LPC	ANN	Indonesian	91.4%
Sonia Sunny et al. [31]	50 speakers, 20 words, one utterance per speaker	LPC	ANN	Malayalam	81.20%
		WPD			87.50%
		DWT			90%
Preeti Saini et al. [32]	9 speakers, 113 words, 3 utterances per speaker	MFCC	HMM	Hindi	96.61%
Noraini Seman et al. [33]	10 speakers, 25 Words, 10 utterances per speaker	MFCC	MLP	Malay	84.73%
Omesh Wadhwani et al. [34]	2 speakers, Discrete words, 120 utterances per speaker	LPC	Fuzzy Neural Network	Hindi Vernacular language	90%

J. 'R. Karam et al. [35]	17 speakers, 11 digits, 2 utterances per speaker	FT	RBF-ANN	NIST English	91.25%
		SCWT			93%
		WPD			94.5%
		DWT			97.8%
K.Daqrouq et al. [36]	27 speakers, 9 vowels, each vowel different no. of utterances	DWT with LPC	PNN	Arabic	93%
Roopa A. Thorat et al. [37]	3 speakers, 3 vowels, 10 utterances per speaker	LPC	Eucledian Distance	Devangiri English	88%
Dr.Yousra F et al. [38]	5 speakers, 7 Words, 2 utterances per speaker	DWT	MLP	Arabic	77%
Vimal Krishnan V.R. et al. [39]	20 speakers, 20 words, 32 utterances per speaker	WPD	MLP	Malayalam	61%
		DWT			89%
T.M. Thasleema et al. [40]	96 speakers, 36 CV units, one utterance per speaker	LPC	ANN	Malayalam	94%
			K-NN		85%
Sherin M. Youssef [41]	30 speakers, 40 words, 3 utterances per speaker	Wavelet Packet Entropy features	ANN	English	Around 96%
Yousef A.A. et al. [42]	17 speakers, 10 digits, 10 utterances per speaker	MFCC	ANN	Arabic	94.5%
			HMM		94.8%
Md. Ali Hossain et al. [43]	10 speakers, 10 digits, 30 utterances per speaker	MFCC	Back propagation NN	Bangala	92%
Gajanan P.K. et al. [44]	4 speakers, 9 words, 1 utterance per speaker	MFCC	Vector Quantization & Euclidean distance	Marathi	88.88%
Ramón Fernández L et al. [45]	72 speakers, 10 digits, 11 utterences per speaker	MFCC	HMM	Spanish	99.67%
			SVM		93.38%
Muhammad G. et al. [46]	100 speakers, 10 digits, 10 utterances per speaker	MFCC	HMM	Bangla	0 to 5, >95% and 6 to -9, <90
Antanas Lipeika et al. [47]	10 speakers, 12 words, 10 utterances per speaker	LPC	DTW	Lithuanian	98.06%
Matthew K. Luka et al. [48]	4 speakers, 10 words, 8 utterances per speaker	MFCC	MLP	Hausa	Best validation .093497

N.S Nehe et al. [49]	100 speakers, 20 words and digits, 20 utterances per speaker	LPCC	Continuous density HMM	Marathi	92.9%
		MFCC			98.2%
		DWT & LPC			99.1%
		WPD & LPC			98.9%
Shady Y. EL-Mashed et al. [50]	20 speakers, 10 separate digits, 50 utterances per speaker	MFCC	SVM	Arabic	94%
Sreejith C et al. [51]	100 speakers, 6 digits, 1 utterance per speaker	MFCC & vector quantization	K-means clustering	Malayalam	88%
Mansour M. Alghamdi1 et al. [52]	1033 speakers, 10 digits, 1 utterance per speaker	MFCC	HMM	Arabic (SAAVB Corpus)	94.13%
Bassam A Q et al. [53]	13 speakers, 33 words, 4 utterances per speaker	MFCC	HMM	Arabic	97.99%
Ling He et al. [54]	15 speakers, 35 words, 1 utterance per speaker	TEO DWT	MLP	SUSAS Database	67.28%
		TEO WPD			82.85%
		TEO PWP			91.56%
		TEO DWT	PNN		78.19%
		TEO WPD			89.45%
		TEO PWP			93.67%
Bishnu Prasad Das et al. [55]	28 speakers, 10 digits, 1 utterance per speaker	LPC	ANN	English	37.5%
		MFCC			51.25%
		LPC, MFCC, ZCR, STE			85%
		LPC	Eucledian distance		23.75%
		MFCC			30%
		LPC, MFCC, ZCR, STE			57.5%
Meysam Mohamad pour et al. [56]	10 speakers, 10 digits, 6 utterances per speaker	Denoising with MFCC & DWT features	MLP & UTA algorithm	Persian	98%
Hemakumar G. et al. [57]	10 speakers, 294 words, 1 utterance per speaker	LPC	Euclidian Distance	Kannada	91.66%
Sukumar A.R. et al. [58]	250 words, isolated question words	DWT	ANN	Malayalam	80%
N. Uma Maheswari et al. [59]	40 speakers, 30 words, once for testing	LPC	Hybrid RBF & ANN	English	91%

Chadawan I et al. [60]	30 speakers, 2 words, 40 utterance per speaker	MFCC	SVM ML classifier	English	SVM performed better
Javed Ashraf et al. [61]	10 speakers, 52 words, 10 utterances per speaker	MFCC	HMM	Urdu	Mean word error rate 5.33%
Cini Kurian et al. [62]	21 speakers, 10 digits, 1 utterace per speaker	PLP Cepstral coefficient	HMM	Malayalam	99.5%
Engin Avci et al. [63]	20 speakers, 25 words, 1 utterance per speaker	WPD	Adaptive Network based fuzzy inference system	Turkish	92%
Md. Akkas Ali et al. [64]	1 speaker, 100 words, 10 utterances per speaker	MFCC	DTW	Bangla	78%
		LPC	DTW		60%
		MFCC & GMM	posterior probability function		84%
		MFCC & LPC	DTW		50%
Malay Kumar et al. [65]	10 speakers, 200 words, 4 utterance per speaker	MFCC	HMM	Hindi	86.08%
		PLP			88.69%
		LPCC			85.21%
		Ensemble ROVER			92.22%
Leena R Mehta et al. [66]	2 speakers, 48 words, 5 utterances per speaker	LPC	Vector quantisation	Marathi	MFCC produced better results
		MFCC			
Mohit Dua et al. [67]	14 speakers , 115 words where no. of words spoken are different for each speaker,	MFCC	HMM	Punjabi	95.63%
Shweta Bansal et al. [68]	2 speakers, 11 words, 2 utterances per speaker	MFCC	DTW	English	70%
A. Akila et al. [69]	2 speakers, 10 words, 4 utterances per speaker	MFCC	HMM	Tamil	90%

From the literature study conducted, it was seen that the recognition accuracy obtained using different languages and methods are different. Moreover, the same feature extraction method produced different results when

used with different classifiers and different languages. So, it is clear that there are many factors that affect the performance of a speech recognition system. Some of the factors include:

a) *Number of speakers* - Recognition accuracy varies with number of speakers.

b) *Standard database* - A standard database and its quality play an important role in the performance of the system.

c) *Language* - Recognition varies across multiple languages.

d) *Age, gender, education, geographical/ cultural differences of speakers* – These factors create variations in the results that are obtained.

e) *Feature extraction method and classifier* – The methods chosen for extracting features and the pattern classifier chosen have a great impact on the performance of the system.

It is therefore difficult to compare the performance of the different methods used for the development of a specific speech recognition system. Moreover, we cannot categorically claim that one particular method is superior to others. The majority of research that has been carried out in speech recognition has contained less number of speakers and the recognition rate was found to be better. Speech is a highly variable signal, characterised by many parameters, and thus large corpora are critical in modelling it well enough for automated systems to achieve proficiency [70].

In the light of the above mentioned issues regarding selection of the best method that is suitable for the overall performance of the speech recognition system, such systems are developed using different techniques and the results obtained are compared to evaluate the performance of these techniques. For developing such systems, selection of the language is an important aspect. There have been numerous works already done in English.

Human computer interaction through conversation in natural language, therefore, plays a very significant role in improving the usage of computers for the common man. Due to the increasing applications of computers, it is necessary to bring human computer interaction as close to human interaction as possible [71]. Malayalam being a language with over 51 alphabets and subtle variations in pronunciations can be considered as an all encompassing work, which can be copied to other languages with relative ease.

Bearing this factor in mind, sixteen different speech recognition systems are developed using different feature extraction techniques and classifier combinations in order to find out the most suitable method for Malayalam. Since speech recognition is a combination of heterogeneous technologies like Signal Processing, Data Mining techniques for Pattern Recognition, Statistical methods and Natural Language Processing techniques, it is essential to have an overall idea of the architecture of the speech recognition system developed. So, section 2.5 discusses the architecture of the different speech recognition systems developed for finding the most suitable combination of the feature extraction method and the pattern classifier, which produces the highest recognition rate for the speech samples.

2.5 Architecture of the Speech Recognition System

For designing an efficient ASR, the primary intention is to design a system whose performance is superior to that of the already existing methods though there exist variability in speech samples [72]. In order to accomplish this objective, three databases are created in Malayalam for this study. Then studies are conducted to compare the performance of these techniques in order to find the optimal combination which is more suitable for this research work. From the literature review, it is seen that most of the works are related to spectral features based on LPC and MFCC and wavelet related methods based on DWT and WPD. So, sixteen different experiments are conducted using a combination of four feature extraction techniques namely LPC, MFCC, DWT and WPD and four classifiers namely ANN, SVM, HMM and Naive Bayes classifiers to select

the feature extraction and classifier combination which produces the best recognition rate.

The architecture of the ASR developed for finding the combination of feature extraction method and the classifier which produces the highest recognition rate is given in figure 2.1 which involves the different stages of developing the speech recognition system.

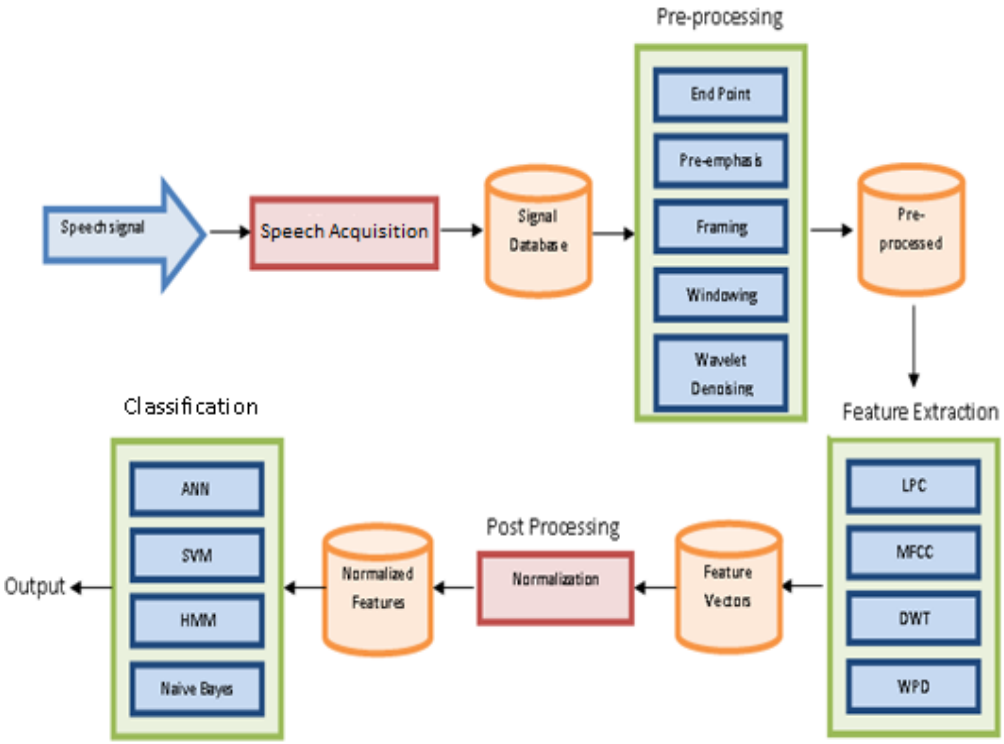


Figure 2.1: Architecture of the ASR for finding the best features and classifier

2.5.1 Creation of the Database

Since there are no built-in standard databases available in Malayalam, three databases namely Vowels database, Digits database and Isolated Words database for this work have been created. The samples stored in the databases are recorded by using a high quality studio-recording microphone at a

sampling rate of 8 KHz (4 KHz band limited). The same configuration and conditions are utilised for the recognition of all the samples in the database. These are stored in the appropriate classes in the database. Speakers from various parts of Kerala were selected to record the speech in order to cover all possible dialectic variations of the language. The samples have been taken from male, female and children of the age group of 6- 70. Obviously, there is bound to be a difference between male and female voices, voices of children, adults and elderly people. This is because of the difference in pitch, frequency, phonetic comprehension and many other factors which occur due to the difference in physiological as well as psychological factors. The speech of children also differ from that of adult males and females.

The speech databases are created using a popular commercial digital audio editing software called GoldWave. It is a highly rated, professional digital audio editor. It allows simplest recording and editing to the most sophisticated audio processing, restoration, enhancements, and conversions. The major features of GoldWave include [73]:

a) *Real-time graphic visuals, such as bar, waveform, spectrogram, spectrum etc.*
b) *filters for noise reduction*
c) *compressor/expander*
d) *volume shaping and matcher*
e) *pitch, reverb and resampling*
f) *ability to support different file formats*
g) *batch processing*
h) *ability to convert a set of files to a different format and apply effects*

i) multiple undo levels and

j) Support for editing multiple and large files at once

The details of the databases created for this research work are given below.

2.5.1.1 Vowels Database

Though the primary objective of this research work is to recognise words in Malayalam, a vowels database is created as a preliminary step in recognising speech since accurate vowel recognition forms the backbone of most successful speech recognition systems. In this work, speech samples are taken from 100 speakers consisting of 40 males, 40 females and 20 children uttering 12 vowels. Thus the database consists of a total of 1200 utterances of the spoken vowels. Though there are 15 vowels in the Malayalam alphabet list, only 12 vowels are taken here. Malayalam has two types of scripts called *pazhaya lipi (old script) and puthiya lipi (new script).* The vowels '**ഌ**' and '**ൠ**' are not included in the new script and the vowel '**ഋ**' is rarely used.

In this study the International Phonetic Alphabet (IPA) format has been used to denote the pronunciation of the words stored in the database. Malayalam has independent vowel letters as well as dependent vowel signs. When a word begins with a vowel sound, an independent vowel letter is used as the first letter of the word. Vowel signs are used word medially and word finally. A consonant letter in Malayalam, it is to be noted, does not represent a pure consonant. It represents a consonant and a short vowel /a/ by default. For instance, the first consonant letter of the Malayalam alphabet is /ka/ and not just /k/. On the other hand, a vowel sign is a diacritic attached to a consonant letter to show that the consonant is followed by a vowel other than /a/. A special diacritic called 'virama' is used to denote a pure consonant sound not followed by a vowel. It is to be noted that the 'virama' is denoted in this study

by an additional vowel symbol /ə/. The vowels of Malayalam are represented in table 2.2 that follows, along with their IPA symbols.

Table 2.2 Vowels and their IPA Format

Vowel	അ	ആ	ഇ	ഈ	ഉ	ഊ	എ	ഏ	ഐ	ഒ	ഓ	ഔ
IPA Format	/a/	/a:/	/i/	/i:/	/u/	/u:/	/e/	/e:/	/ai/	/o/	/o:/	/au/

2.5.1.2 Digits Database

Malayalam digits database consists of the numbers from 0 to 9. A spoken digit recognition process is needed in many applications that need numbers as input—for example, the automated banking system, airline reservations, voice dialing telephone, automatic data entry and the like. For this work, speech samples are taken from 200 speakers. 75 male speakers, 75 female speakers and 50 children are chosen to create the database. This database consists of a total of 2000 utterances of the spoken digits. Table 2.3 shows the Digits in Malayalam, Digits in numeric format, their IPA format and the corresponding English translation.

Table 2.3 Numbers stored in the database in Malayalam, their digit format, IPA format and English translation

Digits in Malayalam	Digits	IPA Format	English Translation
പൂജ്യം	0	/puːʤjam/	Zero
ഒന്ന്	1	/oṉṉə/	One
രണ്ട്	2	/ɾʲaɳɖə/	Two
മൂന്ന്	3	/muːṉṉə/	Three
നാല്	4	/ṉaːlə/	Four
അഞ്ച്	5	/anʤə/	Five
ആറ്	6	/aːrə/	Six
ഏഴ്	7	/eːɻə/	Seven
എട്ട്	8	/eʈʈə/	Eight
ഒൻപത്	9	/onpaɖə/	Nine

2.5.1.3 Isolated Words Database

This is commonly used in command and control applications where the system can recognise a single word command and appropriately respond to the recognised command [21]. Twenty commonly used and meaningful isolated words from Malayalam have been selected to create the database. 1000 speakers including 400 male speakers, 400 female speakers and 200 children were entrusted with the ask of recording the speech samples. Thus the database consists of a total of 20000 utterances of the spoken words. Table 2.4 shows the words in Malayalam, their corresponding words in English, IPA format and the meanings of the words.

Table 2.4 Words stored in the database in Malayalam and English, their IPA Format and meanings in English

Words in Malayalam	Words in English	IPA format	English translation
കേരളം	Keralam	/keːɾʲaɭam/	Kerala
വിദ്യ	Vidya	/ʋiɖʱja/	Knowledge
പൂവ്	Poovu	/puːʋə/	Flower
താമര	Thamara	/t̪aːmaɾʲa/	Lotus
പാവ	Paava	/paːʋa/	Doll
ഗീതം	Geetham	/ɡiːɖam/	Song
പത്രം	Pathram	/pat̪t̪ram/	News paper
ദയ	Daya	/d̪aja/	Mercy
ചിന്ത	Chintha	/ʧint̪a/	Thought
കടൽ	Kadal	/kad̪al/	Sea
ഓണം	Onam	/oːɳam/	Onam
ചിരി	Chiri	/ʧiɾʲi/	Smile
വീട്	Veedu	/ʋiːɖə/	House
□ട്ടി	Kutti	/kut̪t̪i/	Child
മരം	Maram	/maɾʲam/	Tree
മയിൽ	Mayil	/majil/	Peacock
ലോകം	Lokam	/loːkam/	World
മൗനം	Mounam	/maunam/	Silence
വെള്ളം	Vellam	/ʋeɭɭam /	Water
അമ്മ	Amma	/amma/	Mother

ASR is designed to map the acoustic signals captured through a microphone to a set of corresponding words [74]. Designing a speech recognition system involves several independent modules of which the major modules are ***a) front-end processing and b) back end processing*** [75]. During front-end processing, the ***feature extraction*** of the input signal is carried out. During feature extraction, the speech signals are converted to a set of parameters called feature vectors. During back end processing, these features are classified using ***pattern recognition*** techniques to classify them into proper classes. So choosing the best feature extraction method and classification technique play an important role in generating good recognition accuracy.

Though these two modules are the major modules of the speech recognition process, there are two more important modules involved in the effective design of a speech recognition system. They are the ***pre-processing module*** and the ***post processing module.*** Pre-processing techniques are used to tune the speech signals by removing the noise from the signals before extracting features. Post processing techniques are applied to the feature vector set obtained to reduce the dimensions of the feature set obtained and to convert the features to a more consistent and compact format for appropriate classification. One of the problems encountered by ASR is the mismatch between the input and application conditions. The differences that exist in this area can be reduced by these two modules. Complexity and ease of implementation of the speech recognition system mainly depends on these four stages. A brief description of these four stages is given below.

2.5.2 Pre-processing

Pre-processing of speech signals is the first and crucial step in the development of an efficient and robust speech recognition system after creating the database. When voice signals are recorded, different types of degradation components like background noise, noise introduced by environment and recording hardware as well as reverberation and disturbances may interfere with the required speech contents. This affects the quality and clarity of the speech signals and this in turn causes degradation in the performance of the speech recognition system. Moreover, due to the variations in a speech signal, two visually similar waveforms may not produce perceptually similar sounds [2]. So signal pre-processing prior to feature extraction has a great impact on the performance of the speech recognition system and it improves the speech enhancement process.

Different pre-processing algorithms are applied on the captured speech signals to improve their qualities so that they can be made more compatible for further analysis and recognition. In the context of the present research problem, the pre-processing steps applied along with different feature extraction techniques are different. So the detailed pre-processing algorithms and methods are explained in chapter 3 and chapter 4.

2.5.3 Feature Extraction

Feature extraction is the most significant part of speech recognition since it plays an important role in separating one type of speech from other. Choosing relevant features is a crucial step in achieving high recognition performance. In a broad sense, feature extraction can be considered as a data reduction technique since it converts the input signal into a compact set of parameters while preserving the spectral as well as temporal characteristics of

the speech signal information and by discarding the unwanted or redundant information from the signal [76]. Thus feature extraction transforms the signals into a model which can be applied to classification.

Feature extraction produces a stream of vectors that may represent different characteristics depending on the technique used for feature extraction. These characteristics can be spectral characteristics such as Cepstrum, LPC and MFCC features which are obtained over short, overlapping intervals or orthogonal spatial frequency banks using wavelet transforms. So in order to obtain these two types of characteristics, four feature extraction techniques namely *LPC* and *MFCC* based on spectral features and *DWT* and *WPD* based on wavelet features have been adopted in this work. The methodology used for feature extraction and the experiments performed using these techniques and the evaluation of results are illustrated in chapters 3 and 4.

2.5.4 Post Processing

For the proper classification of the feature vectors obtained after feature extraction, they are to be tuned to a more compact form during the post processing stage. The main idea behind post processing is the selection of attribute values that can build a better model than that of taking the attribute values of the entire feature vector set as such. [77]. During this stage, the feature vectors can be converted to a more compact format by bringing their values within a given range. Depending on the dimension of the feature vectors obtained, the higher dimensional feature vector set can be reduced to lower dimension by selecting a subset of the original features retaining the properties of the original data during the post processing stage [78]. The detailed post processing methods used in this work are explained in chapters 3 and 4.

2.5.5 Classification

Speech recognition is one of the core disciplines of pattern recognition. ASR is essentially a multi-class sequential pattern recognition task. Pattern recognition can be viewed as categorising the input data into proper classes via the extraction of significant features or attributes from the data. The complexity of classification depends on the variability in the feature values for samples in the same category relative to the difference between feature values for objects in different classes [79]. Speech recognition involves classification of speech samples into different classes. During the classification stage, training is done using information relating to known patterns. Decisions on classification are made based on the similarity measures from the trained patterns.

Most of the speech classification systems are based on statistical measurements. Statistical learning methods are capable of acquiring knowledge, taking decisions and making predictions from a given set of data. In the present study, various recognition experiments are conducted using different pattern recognition algorithms in order to identify the credibility of the feature parameters obtained. In this work, we have used the well-known approaches that are widely used to solve pattern recognition problems namely *ANN, SVM, HMM and Naive Bayes classifiers* where all these classifiers support multiclass classification and supervised learning.

Since a particular combination of feature extraction method and classifier which is superior to all other methods has not been identified yet, an optimal speech recognition system for Malayalam databases using sixteen experiments with the feature extraction techniques namely LPC, MFCC, DWT and WPD and the pattern classifiers such as ANN, SVM, HMM and Naive Bayes has been located. This also includes the pre-processing and post processing operations.

The results obtained from these experiments are compared and evaluated to locate the system with optimal results. Chapter 3 and 4 elaborate upon the design, development and implementation of these experiments.

2.6 Summary of the Chapter

This chapter is intended to provide an overview of the previous work done in the field of speech recognition. Literature survey has enabled us to see the wide variety of features and classifiers that are used for recognising speech samples. From the study, we can decipher that factors like number of speakers and the quality of the database play an important role in the recognition accuracy. Due to the difficulty in choosing the best feature extraction method and the best classifier which is superior in nature, studies are carried out using different feature extraction methods and classifiers for selecting the best combination which is most suitable for the databases created in Malayalam. It also includes the databases created and the basic architecture which is used for developing the speech recognition system. It is evident that the performance of the speech recognition system results from a combination of several elements, such as versatility of the database used, the strategies used for speech enhancement by removing noise, credibility of the different strategies for feature selection, post processing techniques adopted and the performance of different classifiers and their combinations. So a brief description of these different stages is described in this chapter.

Chapter 3

SPEECH RECOGNITION USING SPECTRAL FEATURE EXTRACTION TECHNIQUES

Contents

This chapter discusses the architecture and implementation of two major feature extraction techniques used in the state-of-the-art ASR systems based on spectral analysis of speech signals namely Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficients (MFCC) along with four classifiers namely ANN, SVM, HMM and Naive Bayes Classifiers. Since there are two feature extraction techniques and four classifiers, a total of eight experiments are carried out on each of the databases created. Experiments are conducted to analyse the effect of the speech recognition system by varying the number of speakers. The implementation results obtained are compared and a performance analysis of these is carried out.

3.1 Introduction

Human beings are capable of identifying and recognising different types of sounds like phonemes (smallest unit of sound), words as well as sentences. However for a machine, it is difficult to differentiate between

different kinds of sounds as human beings perceive it. Suppose a particular word is uttered by different people, the sound waves produced will be different due to the speech variations in individuals. Nevertheless, we are capable of recognising this word because these sound waves will have some common features that humans are able to discern, while in order for a machine to distinguish between the different sounds produced, the important features from the speech signals are to be extracted and made available by the use of different feature extraction techniques.

The principal objective of front end processing in speech recognition is to bring a projection of the speech signal to a feature vector space [15]. From this feature vector space, the relevant information from the speech signals can be extracted easily for further processing. Feature extraction process is considered as the key and most important phase for designing an intelligent and accurate system for the automatic recognition of speech. There are different speech feature extraction techniques available in which the two powerful and dominant techniques which are based on spectral analysis such as Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficients (MFCC) are explained in this chapter which are designed to provide a relevant spectrum of the vocal tract filter. Spectral analysis of speech signals basically involves digital filtering techniques to remove the additive noise and it also emphasises important frequency components of interest. Though a speech signal is non-stationary in nature, it is assumed to be stationary or static during a short period of time [12]. So the speech signal is divided into a number of frames and spectral analysis is done on these frame based segments [80].

The rest of the chapter is organised as follows. Section 3.2 provides an overview of the architecture and the speech recognition process using LPC analysis. The speech recognition process and the architecture of MFCC

analysis are explained in section 3.3. The implementation procedure using both the methods is explained in section 3.4. The performance evaluation of both LPC and MFCC in recognising the speech samples is illustrated in the following section. A comparison of the results obtained using LPC and MFCC are performed in section 3.6. The chapter is concluded in section 3.7.

3.2 Speech Recognition System using LPC

LPC is a digital method for encoding an analog signal. It is based on the mathematical approximation of the vocal tract and models the vocal tract as an Infinite Impulse Response (IIR) system which produces the speech signal. The main characteristics of LPC include [12]:

a) *Encoding of good quality speech at low bit rate*
b) *Good characterisation of the vocal tract*
c) *Estimation of the parameters pitch, formants and spectra and*
d) *Ability to utilise the correlation property of adjacent speech samples in a speech signal.*

LPC assumes that a particular value can be predicted by a linear function of the past values of the signal [13]. Suppose *s(n)* is the speech sample at time *n.* Then it can be expressed using the previous samples *s(n-k)* where *k= 1,...,p*. LPC finds the coefficients $a_{k,}$ which minimises the mean-squared prediction error in terms of the weighted sum of its past samples [12] which can be expressed as

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + ... + a_k s(n-k) \qquad (3.1)$$

In chapter 2, we have seen that after creating the database, there are four stages in developing a speech recognition system namely pre-processing, feature extraction, post processing and classification. So, the speech

recognition system developed using LPC also includes these four stages with a number of computational steps in each stage. Figure 3.1 shows the architecture of the speech recognition system using LPC analysis.

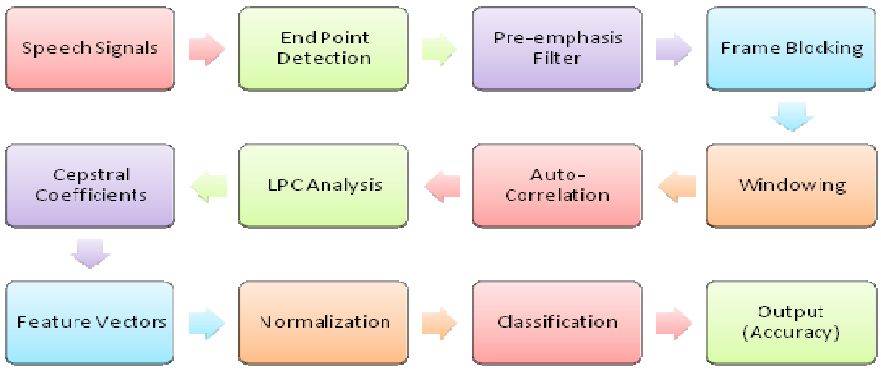


Figure 3.1: Architecture of the speech recognition system using LPC

The different steps involved in the development of a speech recognition system using LPC analysis are given below [12, 30].

3.2.1 Pre-processing

The acoustic signals obtained from the microphone are first pre-processed in order to make it more compatible, noise- free and suitable for feature extraction. In this study using LPC analysis, four techniques are adopted for pre-processing the speech signals namely:

a) *End Point Detection*

b) *Pre-emphasis filtering*

c) *Frame blocking and*

d) *Windowing*

End Point Detection is a common pre-processing technique applied to all types of signals and the other three methods are usually used along with

different feature extraction techniques like LPC and MFCC. A brief description of these techniques is given below.

3.2.1.1. End Point Detection (EPD)

In randomly spoken word recognition systems, there is always a possibility that the spoken word is preceded and succeeded by silence. However adept we are in demarcating the word spoken, there is always a limit to visual editing and so there exists a consequent scope for improvement. Speech consists of voiced and unvoiced parts where the major portion is voiced. Voiced speech is periodic in nature, can be identified and extracted and is the primary ingredient of pre-processing, whereas unvoiced speech is non-periodic and random. So, separating the voiced and unvoiced speech has become a subject of interest and is one of the key pre-processing steps in the speech recognition process [81].

End-pointing techniques can be used to identify word boundaries reliably in isolated word recognition systems. Though it is easy for human beings to locate the beginning and end points of a word, it is a very difficult and complex task for a machine [82]. Since speech signals are time varying signals and are complicated in nature, the signal is sliced into small frames where each frame is considered to be a time-variant signal. There are many algorithms used to find the end points of a speech signal where the most popular methods using time domain parameters to decide the boundary between silence and voice components are:

a) *Zero Crossing Rate (ZCR) and*

b) *Short Time Energy (STE)*

ZCR and STE can be defined as follows.

ZCR: It enumerates the number of times a signal changes its sign from positive to negative and vice versa in a particular frame [83]. Thus ZCR is a measure of the frequency content of the speech signal. Mathematically, this can be expressed as

$$\frac{1}{2.N}.\sum_{m}\left|\operatorname{sgn}\left(s\left(m\right)\right)-\operatorname{sgn}\left(s\left(m-1\right)\right)\right| \tag{3.2}$$

where $1 \le n \le N$ and $1 \le m \le M$

Here, *s(m)* is the speech signal, *M* is the number of samples per frame and N is the number of frames in the signal. If the ZCR is high, it is considered as *unvoiced* and if ZCR is low , it is taken as *voiced* speech.

STE: Energy measures the amount of signal present at a time. The energy of a speech segment is higher than that of a non speech segment [84]. Energy is calculated as

$$E_n = \frac{1}{N}.\sum_{m} s\left(m\right)^2 \tag{3.3}$$

where $1 \le n \le N$ and $1 \le m \le M$

When the energy and zero-crossings are at certain levels, it is assumed to be a speech segment. A proper estimation of the end points is very essential for the proper recognition of isolated words since it avoids the wastage of ASR calculations on the preceding and ensuing silence which in turn reduces the processing complexity [85]. In this study, ZCR is used to find the end points of the speech signal. An example for EPD which is performed in this work is given in figure 3.2. The figure shows the original recorded signal of word 'thamara' and the signal after EPD.

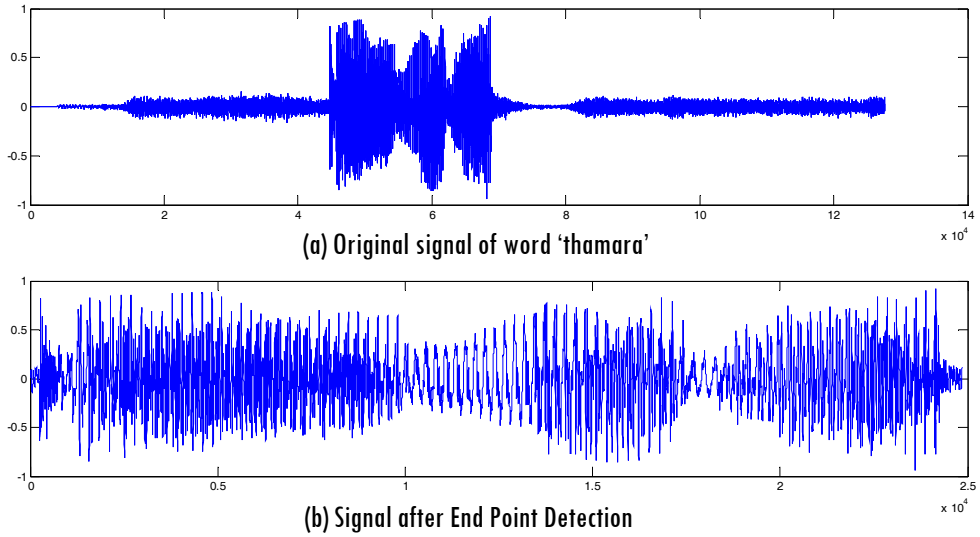


Figure 3.2: (a) Original signal (b) signal after End Point Detection of word 'thamara'

The same procedure is applied to all the signals for finding the start and end points.

3.2.1.2 Pre-emphasis Filtering

A speech signal suffers from additive noise. So, the speech signal is passed through a filter called pre-emphasis filter which emphasises higher frequencies and flattens the speech spectrum by removing the spectral tilt [30]. This filter boosts up the high frequency components of human voice and attenuates the low frequency components of human voice. Pre-emphasis can be denoted as

$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1) \qquad (3.4)$$

Where *s(n)* is the digitised speech sample, *s(n-1)* is the previous digitised speech sample, $\tilde{a}$ is the scaling factor = 0.95, $\tilde{s}(n)$ is the pre-emphasised speech sample and *n* is the number of samples in the whole frame [29].

The acoustic signals obtained from the previous step after performing EPD using ZCR are then applied to the pre-emphasis filter. So, from the above example, the speech signal of word 'thamara' after performing EPD is applied to the pre-emphasis filter. Figure 3.3 demonstrates the speech signal of the word 'thamara' before and the after pre-emphasising.

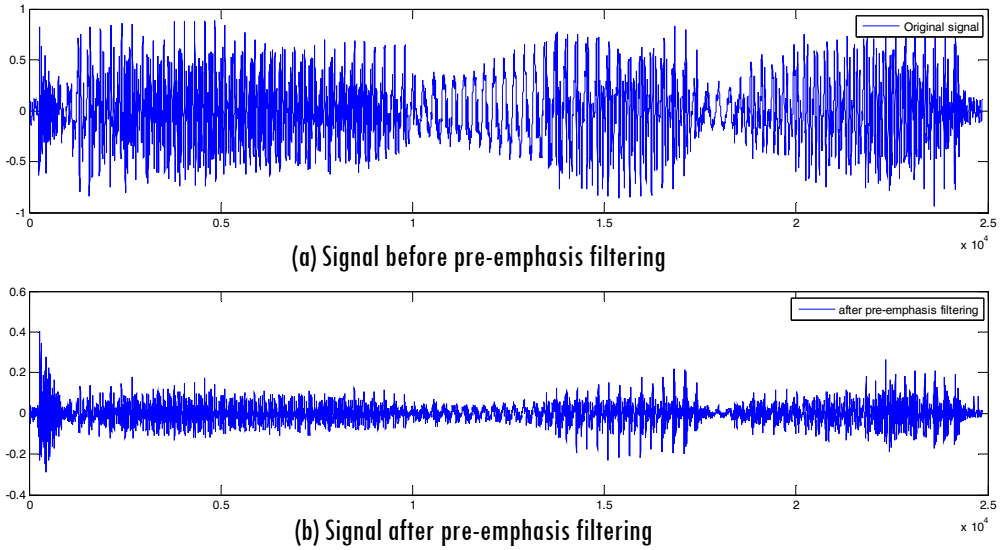


Figure 3.3: (a) Speech signal before pre-emphasis filtering (b) signal after pre-emphasis filtering

3.2.1.3 Frame Blocking

Frame blocking and windowing can be considered as a part of the pre-processing stage or as the preliminary steps in the feature extraction process. During frame blocking, the speech signal is decomposed into a series of overlapping frames where each frame can be analysed independently [12, 37]. Usually, the input speech signal is segmented into frames of 20~30 ms with optional overlap of 1/3~1/2 of the frame size. Each frame is assumed to be stationary. Frame blocking can be expressed as

$$x_l(n) = \tilde{s}(Ml + n) \tag{3.5}$$

where, 0<=n<=N-1 and 0<=l<=L-1

Here, x_l is the l^{th} frame of speech, L is the number of frames within the entire speech signal, N is the total number of samples in the frame, M is the total number of sample spacing between the frames used to measure the overlap.

3.2.1.4 Windowing

Here each sample is multiplied by an N sample window $w(n)$ where the window chosen is the hamming window [29]. It reduces the discontinuities of the speech signal at the edges of each frame which in turn minimises the adverse effects of chopping N samples. Windowing can be expressed as

$$\tilde{x}_l(n) = x_l(n)w(n), \qquad 0 \le n \le N-1 \tag{3.6}$$

The hamming window is defined as

$$w(n) = 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right], \qquad 0 \le n \le N-1 \tag{3.7}$$

3.2.2 Feature Extraction using LPC

The pre-processed signals are then given to the feature extraction procedure which is concerned with the physical analysis of the speech signal. The steps performed in the feature extraction process are:

3.2.2.1 Autocorrelation Analysis

Each frame of the windowed signal is then applied to auto correlation analysis technique which is represented as

$$r_l(m) = \sum_{n=0}^{N-1-m} \tilde{x}_l(n)\tilde{x}_l(n+m), \qquad m = 0,1,...,p \tag{3.8}$$

where, p is the highest autocorrelation value that is taken as the order of the LPC analysis [12] and $r_l(m)$ is the m^{th} autocorrelation of the l^{th} frame.

3.2.2.2 LPC Analysis

During this step, each frame of p + 1 autocorrelations are converted into LPC parameter set by using Levinson Durbin's method [12, 47]. This can formally be given using the following steps.

$$E^{(0)} = r(0) \tag{3.9}$$

$$k_i = \frac{r(i) - \sum_{j=1}^{L-1} a_j^{(i-1)} r(|i-j|)}{E^{(i-1)}}, \qquad 1 \le i \le p \tag{3.10}$$

$$a_i^{(i)} = k_i \tag{3.11}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}, \qquad 1 \le j \le i-1 \tag{3.12}$$

$$E^{(i)} = \left(1 - k_i^2\right) E^{(i-1)} \tag{3.13}$$

The equations from 3.9 to 3.13 are solved recursively for *i= 1,2,.....p.* Here, $a_m(p)$ where $1 \le m \le p$ represents the LPC coefficients and k_m is the Partial Correlation Coefficients (PARCOR coefficients).

3.2.2.3 Conversion to Cepstral Coefficients

The LPC Cepstral coefficients are a very important LPC parameter set which are directly derived from the LPC coefficient set and these features are used as the input data to the classifiers [12, 30]. The recursion used for finding the LPC Cepstral coefficients *c(m)* can be expressed as:

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) . c_k . a_{m-k}, \quad 1 \le m \le p \tag{3.14}$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) . c_k . a_{m-k}, \quad m > p \tag{3.15}$$

3.2.3 Post Processing

The feature vectors obtained are then given to the post processing stage for transforming the feature vector set obtained to a more suitable and appropriate format for the better classification of the features [30]. In this work, the technique used is normalisation.

3.2.3.1 Normalisation

Before using the feature vectors as inputs to the classifiers, the feature vectors are normalised in order to have similar distances between them. The feature vectors obtained are independently normalised by each feature component to a particular range. Here, the feature component x is transformed to a random variable with zero mean and unit variance [86]. This can be represented as

$$\tilde{x} = \frac{x - \mu}{\sigma} \tag{3.16}$$

Where μ and σ are the sample mean and the sample standard deviation of that feature.

3.2.4. Classification

Speech recognition is a fundamental speech classification problem which is applied to the speech features obtained from different feature extraction techniques. During classification, the feature space is partitioned into regions where one region is assigned for each category or class of the input [87]. This is another important stage in the speech recognition process because the speech data are classified into corresponding classes during this stage and is based on supervised learning [88]. Four different classifiers are used here for the efficient recognition of the speech feature set into appropriate

classes namely ANN, SVM, HMM and Naive Bayes classifiers. A brief description of these classifiers are given below.

3.2.4.1 Artificial Neural Networks (ANN)

ANN is a mathematical computational model which is designed to mimic the human brain and is presently used as a very popular and efficient tool in pattern recognition and prediction problems [89]. A Neural Network (NN) consists of a number of interconnected processors called neurons. There are weights associated with the neurons and these are multiplied with the signal value passing through it [90]. ANN is considered to have a remarkable ability in recognising patterns due to its characteristics like:

a) *Adaptive learning*
b) *Parallel organisation*
c) *Fault tolerance and*
d) *Robustness*.

Usually pattern classification operates in one of the following two classification strategies.

- **Supervised classification:** It provides the network with a set of inputs and compares the output with the expected or target values.
- **Unsupervised classification:** It is a more complex and difficult classification method since the target output is not known and the training procedure needs self learning and self organisation [90].

For training the feature vectors using ANN, the first step is to initialise a set of parameters that affect and influence the network learning process. This includes the network topology adopted and the learning parameters chosen such as learning rate and momentum. Network topology provides the structure of the network which deals with the suitable number of hidden layers and the

number of neurons chosen in the hidden layers which in turn improves the mapping between the input and output nodes [29]. The number of neurons in the hidden layer has direct impact on the performance of the ANN which in turn affects the overall recognition rate. If the number of neurons is more, then it may cause over fitting problems and if the number of neurons is too low, it may cause under fitting.

During the training phase, the parameters of ANN vary over time and the network can be monitored and modified for learning purpose [91]. Multilayer Feed Forward Network using Back Propagation algorithm is the most popular NN which is used worldwide in many different types of applications and also in speech recognition [92]. In this work, the ***Multi Layer Perceptron (MLP)*** structure of the ANN which is a feed forward network consisting of multiple layers with one *input layer* which accepts the *N* inputs through *N* parallel input connections, one or more *hidden layers* which accept the weighted sum of the output from the input units and an *output layer* which accepts the weighted sum of the output from the hidden units which finally forms the output [93] is used.

MLP is a supervised learning network as well as a fully connected network [94]. In this work, MLP structure with ***error back propagation algorithm*** [91, 94] is used in which the errors are propagated backwards from the output nodes to the input nodes. The main problem here is to classify the speech sample feature vectors into several speech classes. The number of nodes in the input layer equals the feature dimension whereas the number of nodes in output layer is the same as the number of words in the database [90]. Usually, one hidden layer is enough for efficient classification. The number of nodes in the hidden layer is adjusted empirically for the superior performance of the system. An activation function is applied to the net input to calculate the

output response of a neuron. The network which is adopted in this work uses the ***sigmoid activation function*** where the output varies continuously but not in a linear manner as the input changes [95]. The sigmoid function is expressed as

$$S(x) = \frac{1}{1+e^{-x}} \tag{3.17}$$

where x is the net input.

The training time of a back propagation network increases if the decision regions or desired mappings are complex because as the complexity increases, many repeated execution of the entire training data is necessary to converge [96].

3.2.4.2 Support Vector Machines (SVM)

SVM is one of the powerful classifiers which are used in pattern recognition that uses linear and nonlinear hyper-planes for classifying data. It is basically a binary nonlinear classifier capable of guessing whether an input vector x belongs to class 1 or class 2. For a given set of separable data, the goal is to find the optimal decision function. This is done by choosing a maximum margin as the distance between the closest sample and the decision boundary. It performs classification methods by constructing hyper planes in a multidimensional space that separates different class labels based on statistical learning theory [97, 98]. Now-a-days, SVMs are applied in various fields due to the features of SVM like:

a) *High accuracy and flexibility*
b) *Capacity to accommodate large number of attributes*
c) *Ease of training and*
d) *Ability to model complex and real-world problems.*

SVM is a kernel-based algorithm. A kernel is a function that transforms the input training data from the input space to a higher-dimensional feature space and then separates these data in the new space. Kernel functions can be linear or nonlinear [99]. Kernels can extend the decision boundaries set by SVM to non-linear boundaries. The most commonly used kernels are *a) Simple linear kernel, b) polynomial kernel, c) Gaussian kernel (Radial Basis Function) RBF kernel and d) Sigmoid kernel* [97, 99]. In this research work, the kernel used is *polynomial kernel.* It is a non-stochastic kernel and popular method for non linear modelling which provides good classification accuracy with minimum number of support vectors and low classification error.

Multi-class Classification using SVM: Though SVM is inherently a successful and popular binary nonlinear classifier, it is extended to multiclass classification since ASR is a multiclass problem [100]. In binary classification, only the decision boundaries of one class are to be considered and the rest is considered as second class. But in multiclass classification, several boundaries are essential and so it is a more complex task than binary classification. There is also a probability for more errors to occur. In order to deal with multi class problems, the conventional way used is to decompose the *M*-class problem into a series of two-class problems and construct several binary classifiers where, *M* denotes the number of classes. There are two major strategies for multiclass classification namely *One-against-All* and *One-against-One* or *Pair wise* classification.

***One-against-All*:** A classification problem is divided into *K* binary problems to classify among *K* classes, where each problem separates a given class from the other $K-1$ classes [97, 101]. This requires $N = K$ binary classifiers. The i^{th} SVM will be trained with all of the examples in the i^{th} class with positive labels, and all other examples with negative labels. SVMs trained in this way

are also referred as *one-versus-rest* [102]. The final output of this is the class that corresponds to the SVM with the highest output value. Here, the optimal hyper plane that separates each class from the rest of the classes is found out. A drawback of this method is that there is no bound on the generalisation error for this and the training time of the standard method scales linearly with *N*.

***One–against-One*:** In this approach, each class is compared to the other class and a binary classifier is built to discriminate between each pair of classes, while discarding the rest. Then it constructs all possible two-class classifiers from a training set of *N* classes, where each classifier is trained on only two out of *N* classes [96]. There is one binary SVM for each pair of classes to separate members of one class from those of the other [101]. Using this method, the whole system can be trained with a maximum number of different samples for each class, with a limited computer memory [102]. This requires building of *K(K−1)/2* binary classifiers and is considered to have a smaller dimension.

3.2.4.3 Naive Bayes Classifiers

Naive Bayes classifiers are based on the Bayesian theory presented in 1973 which is a simple and effective probability classification method. This is also a supervised classification technique. For each class value it is estimated that a given instance belongs to that class [103]. The feature items in one class are assumed to be independent of other attribute values called class conditional independence [104]. Naive Bayes classifier needs only a small degree of training set to estimate the parameters for classification and this model uses the maximum likelihood technique for parameter estimation in most of the applications. The classifier can be stated using the equation

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{3.18}$$

where

- P(A) is the prior probability or the marginal probability of A, which is independent of B.
- P(A|B) is the conditional probability of A, given B called the posterior probability. It is derived from or dependent on the value of B.
- P(B|A) is the conditional probability of B given A.
- P(B) is the prior or marginal probability of B which acts as a normalising constant.

The probability value of the winning class dominates over that of the others [105]. Naive Bayes classifiers are successfully applied in different areas like medical diagnosis, document classification and parameter estimation due to

a) *its ease in the simple implementation and interpretation process of the algorithm*
b) *attribute independence and*
c) *fast and efficient training procedure.*

3.2.4.4 Hidden Markov Models (HMM)

HMM is the most frequently employed core technique and a very successful pattern recognition method used in the field of speech recognition. It is used in acoustic modelling as well as statistical signal processing. It is a mathematical model derived from Markov models. HMM is a stochastic signal model which consists of a number of probabilistic states and state transitions which represent the model change from one current state to another. The states are interpreted as acoustic models and the transitions provide temporal constraints indicating how the states follow each other in a sequence [13]. In

speech recognition applications, since speech always goes forward with time, transitions also move forward or may make a self loop [12]. HMM consists of a number of elements. The different parameters in HMM include [106, 107].

- N - Number of states in the model
- S = $\{s_1, s_2, ..., s_N\}$ – Set of all possible states in the model
- M - The number of distinct observation symbols per state.
- V = $\{v_1, v_2, ..., v_M)$ – Individual symbols
- Q = $\{q_1, q_2, ..., q_t\}$ states at a particular time
- O = $\{o_1, o_2, ..., o_t\}$ possible observation sequence
- Each transition in the state diagram of an HMM is associated with a transition probability denoted by matrix A. The state transition probability storing the probability of state *j* following state *i* can be expressed as

  A = $\{a_{ij}\}$ where,

$$a_{ij} = P\,(q_{t+1} = s_j \mid q_t = s_i), \quad (1 \le i, j \le N) \tag{3.19}$$

- Each state is associated with a set of discrete symbols with an observation probability assigned to each symbol denoted by B. This can be represented as B = $\{b_j(k)\}$, in which

$$bj(k) = P(O_t = v_k \mid q_t = s_j), \quad 1 \le k \le M \tag{3.20}$$

- Initial state distribution $\pi = \{\pi_i\}$, in which

$$\pi_i = P(q_1 = s_i), \quad 1 \le i \le N \tag{3.21}$$

In a compact form, an HMM can be defined as a triplet,

$$\lambda = (A, B, \pi) \tag{3.22}$$

to indicate the complete parameter set of the model. HMM is a statistical model which is assumed to be in a Markov process with unknown parameters [12]. HMM can be considered as the simplest dynamic Bayesian network because the main challenge of HMM is to find all the appropriate hidden parameters from the observable states [108]. In a regular Markov model, the state is directly visible to the observer but in an HMM, the state is not directly visible. Though the state is not visible, the variables influenced by the states are visible. Each transition in the state diagram of a HMM has transition probability associated with it [106].

3.3 Speech Recognition System using MFCC

MFCC is another popular spectral analysis method which is a special case of homomorphic signal processing. MFCC can be defined as the Discrete Cosine Transforms (DCT) of the log filter bank amplitudes.

Figure 3.4 depicts the detailed schematic diagram of the speech recognition procedure using MFCC.

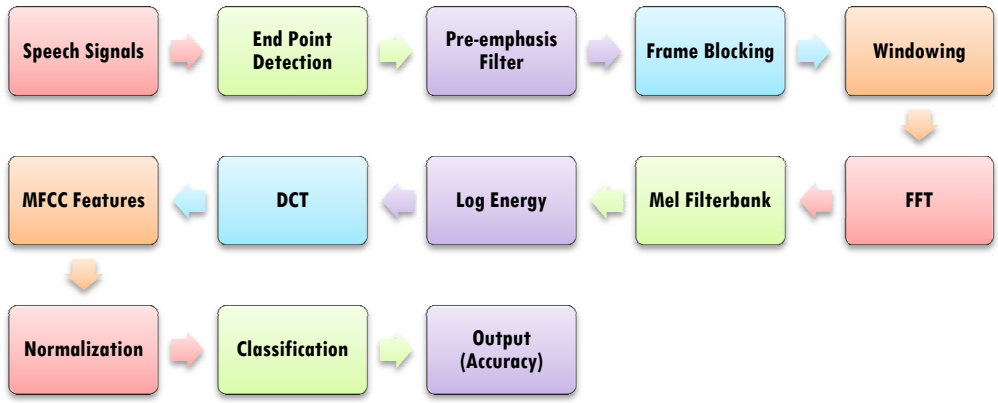


Figure 3.4: Schematic diagram of the speech recognition system using MFCC

Some of the important characteristics of MFCC are:

a) *They are less dependent on speaker-dependent characteristics and*

b) *They improve recognition in noisy environments.*

Since MFCC and LPC are based on spectral analysis of the speech signals, the same pre-processing steps namely ***a) End point detection b) pre-emphasis filtering c) frame blocking and d) windowing,*** post processing method namely ***normalisation*** and pattern classifiers such as ***a) ANN, b) SVM, c) HMM and d) Naive Bayes classifiers*** which were used for LPC analysis were also employed here. So this section discusses only the feature extraction procedure using MFCC.

3.3.1 Feature Extraction using MFCC

After pre-processing the speech signals, the signals are subjected to the feature extraction procedure using MFCC for extracting features. This consists of four computational steps and the steps in the feature extraction process are given below.

3.3.1.1 Fast Fourier Transform

Since the speech signals correspond to different energy distribution over frequencies, compute the Fast Fourier Transform (FFT) of each frame to obtain the magnitude frequency response of each frame [33]. FFT computation can be expressed as

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-j2\pi nk}{N}}, 0 \le k < N \tag{3.23}$$

Where *x[n]* is the windowed frame, *X[k]* is the Discrete Fourier Transforms (DFT) of the frame which represents the magnitude and phase of that frequency component in the original signal [42].

3.3.1.2 Mel Filter Bank

Human hearing is not equally sensitive to all frequency bands. Human perception of frequency is less sensitive at higher frequencies and is non-linear in nature. The perceived Mel scale frequency f_{mel} can be computed from the real linear frequency of the speech signal f_{lin} as [15]

$$f_{mel} = 2595 * LOG_{10}\left[1 + \frac{f_{lin}}{700}\right] \tag{3.24}$$

So during the next step, the magnitude spectrum *X[k]* is passed through the mel filter bank by multiplying each FFT magnitude coefficient by the corresponding filter value [43]. A filter bank consists of a set of band pass filters whose bandwidths and spacing are almost equal to those of critical bands. Moreover, the range of the centre frequencies of these filters cover the most important frequencies for speech perception [109]. If there are *M* filters in the filter bank, a set of *M* values represents the energy in each band.

3.3.1.3 Log Energy

Logarithm compresses dynamic range of values. Human response to signal level is logarithmic because humans are less sensitive to slight differences in amplitude at high amplitudes than low amplitudes [45]. So compute the log energy at the output of each filter which is represented as

$$s[m] = In\left[\sum_{k=0}^{N-1}\left[X[k]\right]^2 H_m[k]\right], 0 \le m < M \tag{3.25}$$

3.3.1.4 Discrete Cosine Transforms

During the final step, the log mel spectrum is converted back to time which are the Mel Frequency Cepstral Coefficients (MFCC) [46]. The cepstral representation of the speech spectrum provides a good representation of the

local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT) [48]. This can be specified as

$$c[n] = \sum_{m=0}^{M-1} s[m] \cos\left(\frac{\pi n (m - 0.5)}{M} \right). \tag{3.26}$$

3.4 Implementation

The Malayalam dataset used for implementation consists of 100 speakers uttering 12 vowels, 200 speakers uttering 10 digits and 1000 speakers uttering 20 isolated words. The speech recognition systems are implemented using two powerful softwares namely MATLAB and WEKA.

MATLAB (Matrix Laboratory): The feature extraction, pre-processing and post processing steps of the speech recognition system design are implemented using MATLAB which is a high-level language and an interactive environment for numeric and scientific computations, visualisation, application development and programming. It is a software package which has different tool boxes and a variety of functions for analysing data, developing algorithms, and for creating different types of models and applications [110]. It is a flexible and high performance language for technical computing and a useful tool which is used in many application areas since it provides an interactive environment for design and problem solving.

The main features of MATLAB [111] include:

a) *Mathematical functions for Linear Algebra, Statistics, Fourier Analysis, filtering, optimisation, numerical integration, and solving ordinary differential equations.*

b) *Built-in graphics for visualising data and tools for creating custom plots.*

c) *Different tools for improving code quality and maintainability and maximising performance*

d) *Building applications with custom graphical interfaces and*

e) *Ability to integrate the algorithms developed with external applications and languages such as C, Java, .NET, and Microsoft Excel.*

Though MATLAB has a number of toolboxes for different applications, the main toolbox exploited in this work is the signal processing toolbox. It consists of a rich set of built-in functions and algorithms for the efficient processing of the speech signals. Some of the common features include signal and linear system models, signal transforms like FFT, DFT, STFT, wavelets, waveform and pulse generation functions, statistical signal measurements and data windowing functions, digital FIR and IIR filter design, analysis and implementation methods, linear prediction and parametric time-series modelling. This also includes analysis and visualisation tools for verifying accuracy and performance of the systems developed.

WEKA (Waikato Environment for Knowledge Analysis): The classification part is implemented using a package called WEKA, which is an open-source software with a collection of machine learning algorithms developed in JAVA for data mining tasks [112]. These algorithms can be directly applied to the feature vector set obtained from feature extraction. WEKA is very versatile and flexible software which also can implement algorithms for data pre-processing, classification, regression, clustering and association rules as well as visualisation tools. WEKA has a number of interfaces where WEKA Explorer is the one

which is used to perform different tasks during classification and analysis. It has a set of panels, where each of them can be used to perform a certain task.

The data file format normally used by WEKA is Attribute Relation File Format (ARFF) or Comma Separated Values (CSV) format. This work utilises the ARFF file format which consists of special tags like attribute names, attribute types, attribute values and the data for representing the features in the data file. In this work, the k-fold cross-validation technique is used where the original sample is randomly partitioned into k subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k - 1$ subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The k results from the folds then can be averaged or combined to produce a single estimation.

The main advantages of WEKA are:

a) *It provides many different algorithms for data mining and machine learning*

b) *It is an open source and is freely available*

c) *It is platform- independent and*

d) *It is easy to use.*

3.4.1 Implementation of Speech Recognition System using LPC

As explained above, the implementation procedure for extracting features using LPC analysis involves different computational steps. The algorithm for the feature extraction using LPC is given in table 3.1.

Table 3.1: Algorithm for feature extraction using LPC analysis

1. For each speech sample perform the following steps:

1.1 Find the start and end point of the signal using the equations of ZCR given in 3.2.

1.2 Apply pre-emphasis filtering to the digitised speech signal using the eqn. 3.4.

1.3 Perform frame blocking by dividing the speech samples into 30 ms window frames using the eqn. 3.5.

1.4 Find the number of samples in the frame using the following calculation

Number of samples in the frame= Sampling rate frame length*

*8000 samples/sec * 0.030 seconds = 240 samples.*

1.5 Separate adjacent window frames by 80 samples with 160 overlapping samples.

1.6 Multiply each frame with a hamming window function which generates 240 discrete points for the hamming window using eqn. 3.6.

1.7 Calculate the auto correlation coefficients of each window frame of 240 samples using the eqn.3.8.

1.8 Apply Levinson Durbin algorithm to the auto correlation coefficients using the eqns. 3.9 to 3.13.

1.9 Convert the auto correlation coefficients to Cepstral coefficients which produces LPC coefficients of order 10 using the eqns. 3.14 and 3.15.

3.4.2 Implementation of Speech Recognition System using MFCC

The implementation procedure using MFCC also involves different computational steps. The algorithm for feature extraction using MFCC is given in table 3.2.

Table 3.2: Algorithm for feature extraction using MFCC analysis

1. For each speech signal perform the following steps.

1.1 Find the start and end point of the signal using endpoint detection technique given in eqn. 3.2.

1.2 Pass each speech sample through a filter which emphasises the signal given by the equation 3.4.

1.3 Segment the speech samples into frames of length 20 ms using the equation 3.5.

1.4 Calculate the number of samples in the frame using the following calculation

Number of samples in the frame= Sampling rate frame length*

*8000 samples/sec * 0.020 seconds = 160 samples.*

Set the overlap between successive frames at 80 samples.

1.5 Multiply each frame by a hamming window where hamming window is defined as in equation 3.6

1.6 Apply FFT to each windowed segment of speech to convert each frame of 160 samples from time domain to frequency domain using the equation 3.23.

1.7 Filter the power spectrum obtained by a series of overlapping filters that are centered on the Mel scale.

1.8 Take the log energy at the output of each filter using the equation 3.25.

1.9 Compute the DCT to convert the log mel spectrum back to time domain as given in equation 3.26.

1.10 Extract an output of Mel Cepstral coefficients of 13^{th} order where the first 12 are used and the 13^{th} one is the energy coefficient.

3.5 Performance Evaluation

As mentioned above, a total of eight experiments are conducted in order to evaluate the performance of LPC and MFCC along with different classifiers in recognising speech. There are different methods for evaluating the performance of a speech recognition system. The major metrics that are used for the performance evaluation are Recognition Accuracy and Confusion Matrix.

Recognition Accuracy: The main criterion for measuring the performance of a speech recognition system is the recognition accuracy, which is a practical value and an important measure for all speech recognition applications. Improving the accuracy of an ASR is one of the most important research challenges after years of research and development [113]. It is measured as the number of utterances recognised correctly out of the total number of utterances spoken which is expressed as a percentage. If we have two classes, then we can refer in terms of positive tuples and negative tuples – ie *True Positive, False Positive, True Negative and False Negative*. In this classification problem, the aim is to classify a test data into one of the twelve distinct categories for vowels, ten distinct categories for digits and twenty distinct categories for isolated words. The following crieteria is used for claculating the accuracy.

Correctly Selected $\rightarrow$ True Positive

Mistakenly Selected $\rightarrow$ False Positive

Correctly Rejected $\rightarrow$ True Negative

Mistakenly Rejected $\rightarrow$ False Negative

A good classification test always results in high values for accuracy. The overall recognition rate in terms of accuracy can be computed as below:

$$Accuracy = \frac{\left(True\ Positive + True\ Negative\right)}{\left(True\ Positive + False\ Positive + True\ Negative + False\ Negative\right)} \quad (3.27)$$

Recognition accuracy is represented as percentage.

Confusion Matrix: A confusion matrix is a performance analysis tool which is used to represent the results in a matrix format. The diagonal elements of a confusion matrix contain the instances that are correctly classified. Recognition accuracy provides only the percentage of correctly and wrongly classified instances. But a confusion matrix provides more information about where the classifier failed in recognising the features and detailed class-conditional error rates.

Confusion Matrix is a useful tool for analysing how well the classifier can recognise tuples of different classes. If there are *m* classes, a confusion matrix table will be of at least size *m* by *m*. For a classifier to have good accuracy the tuples along the diagonal of the confusion matrix should contain large values and the rest of the entries should have zero or very small values. The table may have additional rows or columns for representing totals or recognition rates per class. A confusion matrix can be represented using the equation

$$Eji = \Pr\{decision \quad j \mid class \quad i\} \tag{3.28}$$

which denotes the matrix of counts where the true class *i* is classified as *j* [88].

3.5.1 Performance Evaluation of Speech Recognition System using LPC Analysis

After pre-processing and feature extraction, the feature vector set obtained are normalised using the eqn. 3.16 and are given for classification. The experiments are done by varying the number of speakers to evaluate the impact of the number of speakers in the recognition rate because when the number of speakers increases, then the recognition rate decreases. In the literature study, we

have seen that small vocabulary databases with only few numbers of speakers produce high accuracy. Table 3.3, table 3.4 and table 3.5 show the results obtained after classification using the four classifiers on the three datasets.

Table 3.3: Classification results for LPC using MLP, HMM, Naive Bayes and SVM classifiers on Vowels database

No. of Speakers	Total Samples	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	120	115	95.83	112	93.33	113	94.16	114	95.0
25	300	284	94.66	275	91.66	278	92.66	281	93.66
50	600	560	93.33	540	90.0	547	91.16	554	92.33
75	900	834	92.66	806	89.55	815	90.55	821	91.22
100	1200	1101	91.75	1059	88.25	1072	89.33	1088	90.66

Table 3.4: Classification results for LPC using MLP, HMM, Naive Bayes and SVM classifiers on Digits database

No. Of Speakers	Total Samples	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	100	96	96	93	93	94	94	96	96
25	250	237	94.8	229	91.6	232	92.8	235	94.0
50	500	467	93.4	452	90.4	458	91.6	462	92.4
75	750	694	92.53	671	89.46	680	90.66	685	91.33
100	1000	918	91.8	886	88.6	895	89.5	909	90.9
150	1500	1360	90.66	1301	86.73	1328	88.53	1337	89.13
200	2000	1776	88.8	1718	85.9	1734	86.7	1754	87.7

Table 3.5: Classification results for LPC using MLP, HMM, Naive Bayes and SVM classifiers on Isolated Words database

No. Of Speakers	Total Samples	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	200	195	97.5	190	95.0	192	96	193	96.5
50	1000	958	95.8	927	92.7	936	93.6	948	94.8
100	2000	1863	93.15	1806	90.3	1826	91.3	1851	92.55
250	5000	4502	90.04	4386	87.72	4437	88.74	4472	89.44
500	10000	8664	86.64	8360	83.60	8481	84.81	8580	85.8
750	15000	12502	83.34	11882	79.21	12118	80.78	12310	82.06
1000	20000	15978	79.89	15101	75.5	15325	76.62	15602	78.01

The graphs showing the performance of various classifiers on the three databases based on recognition accuracy and number of speakers using LPC are given in figure 3.5, figure 3.6 and figure 3.7.

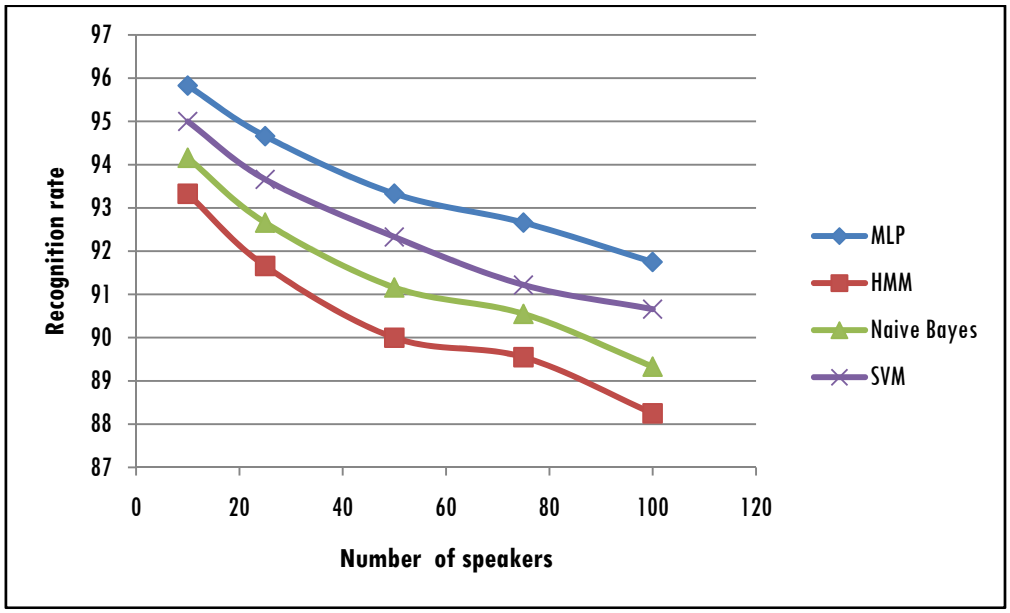


Figure 3.5: Performance of different classifiers on Vowels database using LPC

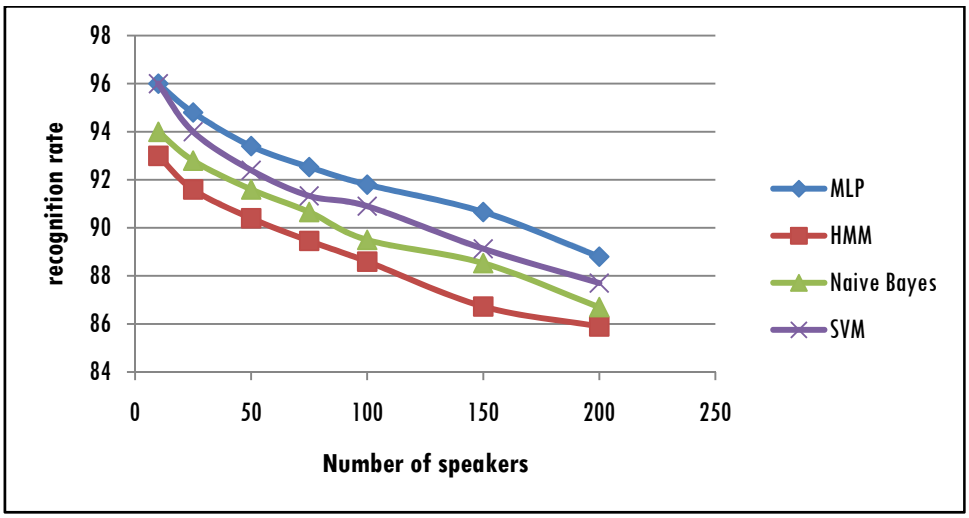


Figure 3.6: Performance of different classifiers on Digits database using LPC

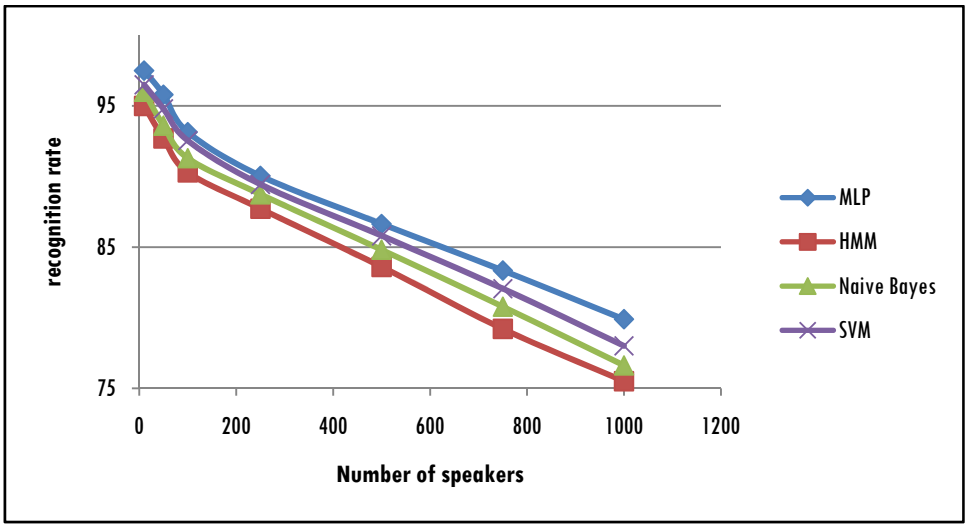


Figure 3.7: Performance of different classifiers on Words database using LPC

From the results obtained using LPC features, it is observed that MLP classifier produced optimal results. Though three databases were created for comparing the performance of various speech recognition systems in this work, the main priority and concern is given for recognising isolated words, which has a sufficient number of speech samples (1000 speakers uttering 20 words each). So the results obtained from the isolated words database alone are taken for further representations.

Since better results are obtained using the MLP structure of the ANN classifier, the results obtained using ANN are selected to represent the confusion matrix. An overall recognition accuracy of 79.89% is obtained for LPC and MLP combination. Figure 3.8 shows the confusion matrix generated for isolated spoken words with 1000 people uttering 20 words.

Predicted Class

Actual Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	687	18	0	1	45	0	4	42	0	9	13	13	35	2	29	43	12	0	15	32
2	1	814	17	11	0	13	27	5	3	26	5	0	12	16	0	10	20	0	3	17
3	0	38	840	0	11	0	2	4	0	3	0	0	4	0	21	2	20	4	24	27
4	1	21	3	745	36	20	28	1	3	11	5	0	15	0	15	3	32	22	16	23
5	0	1	0	29	813	0	10	17	21	0	14	0	20	3	26	0	2	24	0	20
6	3	34	7	1	18	769	4	0	18	4	12	25	1	1	1	26	10	33	22	11
7	0	25	0	12	0	5	862	0	1	27	0	6	2	23	12	15	0	0	10	0
8	24	11	0	8	1	2	1	815	0	26	1	6	0	20	43	0	2	17	0	23
9	0	31	18	4	12	21	48	0	786	10	0	14	2	0	26	5	9	11	2	1
10	4	19	1	2	0	7	1	17	1	820	0	1	0	9	39	14	2	0	1	62
11	1	12	33	2	13	0	3	3	0	14	805	0	0	1	18	1	11	45	5	33
12	25	8	4	2	4	12	9	1	12	3	0	844	9	12	27	28	0	0	0	0
13	16	43	10	1	0	11	1	3	26	1	0	4	781	0	11	35	32	0	4	21
14	1	20	9	1	0	21	8	10	2	22	0	14	0	858	13	0	0	20	0	1
15	20	12	11	0	1	10	3	30	0	28	1	0	10	0	840	11	0	22	1	0
16	26	4	14	0	0	1	0	0	12	18	0	0	2	1	13	879	0	0	28	2
17	2	43	11	9	1	19	1	1	2	1	1	22	20	40	2	4	698	56	46	21
18	1	10	56	11	12	0	2	0	0	2	22	0	1	0	5	2	97	722	37	20
19	14	71	11	1	1	1	0	4	0	1	3	0	33	0	18	35	11	13	759	24
20	5	18	1	6	16	4	0	18	4	24	12	5	0	6	14	0	9	17	0	841

Figure 3.8: Confusion matrix for Isolated Words database using LPC MLP combination

3.5.2 Performance Evaluation of Speech Recognition System using MFCC Analysis

The results obtained after classification using the 4 classifiers on each database based on recognition accuracy, comparison graphs and confusion matrix using MFCC are explained in this section. Table 3.6, table 3.7 and table 3.8 shows the results obtained after classification using the four classifiers on the MFCC feature vector set.

Table 3.6: Classification results for MFCC using MLP, HMM, Naive Bayes and SVM classifiers on Vowels database

No. Of Speakers	Total Samples	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	120	116	96.66	113	94.16	114	95.0	115	95.83
25	300	287	95.66	277	92.33	280	93.33	284	94.66
50	600	565	94.16	549	91.5	553	92.16	560	93.33
75	900	840	93.33	811	90.11	819	91.0	828	92.0
100	1200	1109	92.41	1069	89.08	1080	90.0	1100	91.66

Table 3.7: Classification results for MFCC using MLP, HMM, Naive Bayes and SVM classifiers on Digits database

No. Of Speakers	Total Samples	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	100	97	97.0	94	94.00	95	95.0	96	96.0
25	250	240	96.0	231	92.4	236	94.4	238	95.2
50	500	472	94.4	458	91.6	463	92.6	469	93.8
75	750	702	93.6	680	90.66	687	91.6	697	92.93
100	1000	925	92.5	896	89.6	904	90.4	915	91.5
150	1500	1369	91.26	1324	88.26	1337	89.13	1354	90.26
200	2000	1794	89.7	1739	86.95	1753	87.65	1776	88.8

Table 3.8: Classification results for MFCC using MLP, HMM, Naive Bayes and SVM classifiers on Isolated Words database

No. Of Speakers	Total Samples	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	200	197	98.5	192	96.0	194	97.0	195	97.5
50	1000	972	97.2	948	94.8	955	95.5	963	96.3
100	2000	1913	95.65	1858	92.9	1861	93.05	1890	94.5
250	5000	4604	92.08	4494	89.88	4530	90.6	4565	91.3
500	10000	8861	88.61	8502	85.02	8686	86.86	8775	87.75
750	15000	12734	84.89	12154	81.02	12448	82.98	12575	83.83
1000	20000	16354	81.77	15452	77.26	15696	78.48	16054	80.27

The feature vector set obtained using MFCC analysis are classified using the above four pattern classifiers. Figure 3.9, figure 3.10 and figure 3.11 give a comparison of the results obtained using MFCC analysis and the 4 classifiers on the three databases based on rate of recognition and number of speakers.

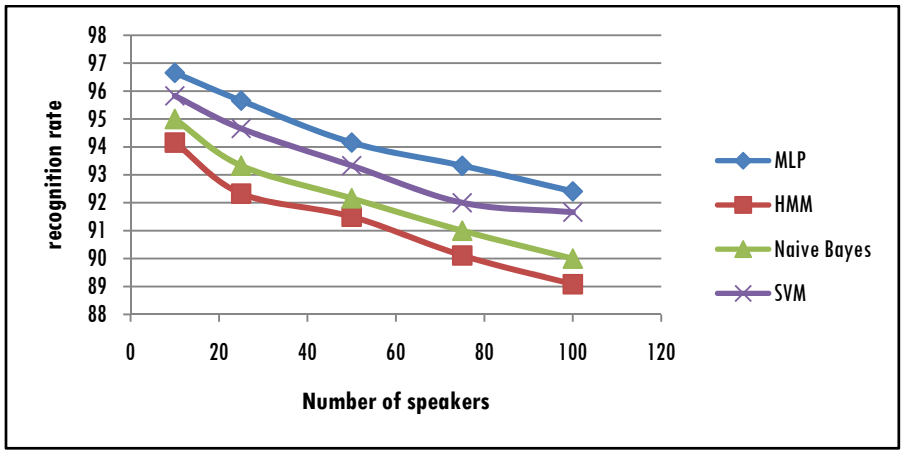


Figure 3.9: Performance of different classifiers on Vowels database using MFCC

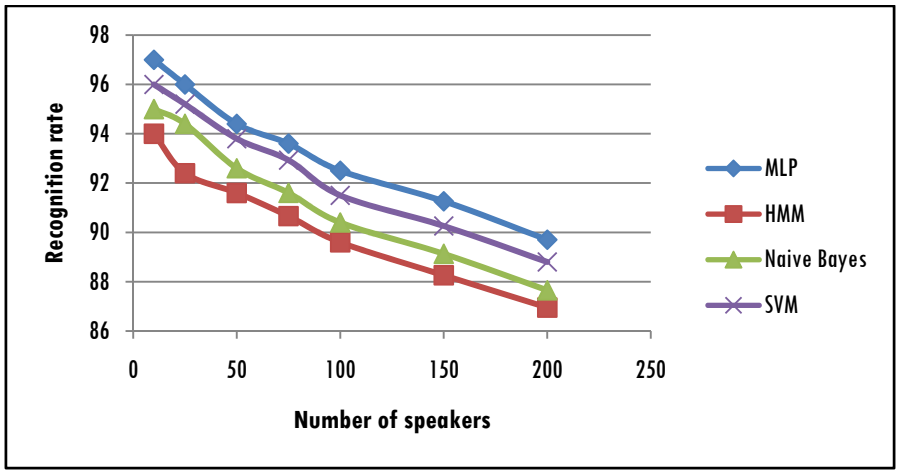


Figure 3.10: Performance of different classifiers on Digits database using MFCC

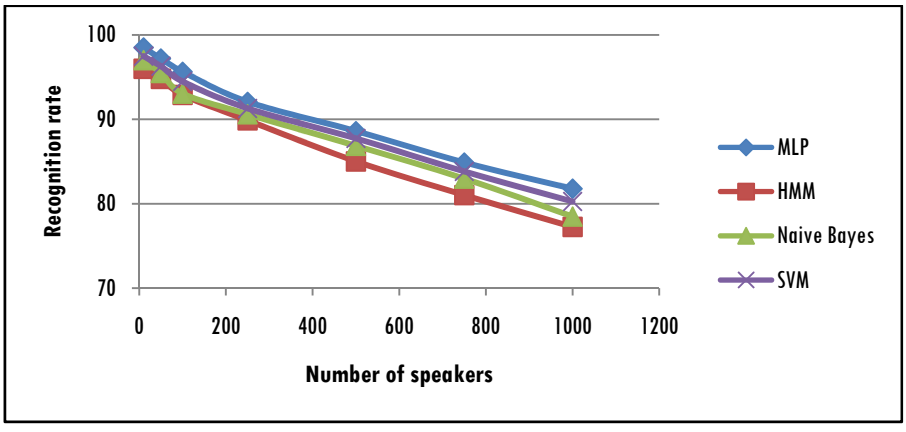


Figure 3.11: Performance of different classifiers on Words database using MFCC

The confusion matrix obtained using MFCC features and MLP classifier on isolated spoken words with 1000 speakers is given in figure 3.12. An overall recognition rate of 81.77% is obtained using the MFCC and MLP combination.

Predicted Class

Actual Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	982	4	0	0	0	0	3	0	0	0	2	0	5	0	2	0	2	0	0	0
2	29	745	0	34	0	35	9	18	20	7	0	32	11	0	13	16	7	7	0	17
3	0	15	830	0	0	13	0	35	17	0	4	0	18	0	4	22	0	17	0	25
4	10	0	16	823	0	4	17	18	11	7	12	4	11	0	13	13	12	9	8	12
5	0	7	0	4	900	0	10	0	15	3	20	0	0	12	0	6	8	0	15	0
6	10	0	0	4	0	950	3	5	0	0	5	0	0	8	6	0	2	0	0	7
7	0	4	0	14	0	6	879	0	7	0	16	14	0	15	0	5	5	2	0	33
8	6	13	0	0	0	0	7	912	0	8	12	0	0	20	4	3	0	5	0	10
9	10	8	20	2	17	22	24	33	697	24	5	8	17	18	10	38	4	20	2	21
10	6	0	1	7	0	3	9	0	5	935	0	13	3	0	5	0	0	4	5	4
11	20	7	12	15	25	23	7	34	26	28	650	25	5	12	32	10	30	13	8	18
12	22	12	23	14	15	25	16	18	17	8	11	720	0	15	13	15	23	8	13	12
13	20	0	10	4	0	14	0	11	0	16	0	0	890	0	13	0	14	0	5	3
14	15	18	8	22	65	10	2	20	3	15	2	30	1	726	1	23	14	6	7	12
15	4	18	24	0	31	0	14	5	18	15	20	11	0	90	677	35	0	6	7	25
16	5	0	0	8	0	4	0	8	2	0	7	0	1	3	0	960	0	1	1	0
17	24	11	15	32	9	25	14	5	12	20	23	12	19	24	0	11	666	15	25	38
18	5	9	2	6	3	4	12	5	7	0	1	0	5	3	15	14	0	899	0	10
19	14	10	6	14	1	5	6	10	2	1	2	7	13	5	9	8	0	2	880	5
20	18	10	26	12	16	50	14	16	35	7	10	4	5	15	12	48	15	45	9	633

Figure 3.12: Confusion matrix for Isolated Words database using MFCC MLP combination

3.6 Comparison of LPC and MFCC Methods

From the results obtained, it is observed that the recognition rate obtained using MFCC is slightly better than that of LPC in recognising the

speech samples for Malayalam. One of the drawbacks of LPC- based speech features is that it tends to include speaker-dependency into modelling, which in turn degrades the performance particularly for a speaker-independent recognition system. Though LPC analysis is widely used, Cepstral Analysis using MFCC provides more robust speech features since they are less dependent on speaker-dependent characteristics and improves recognition in noisy environments.

For all the three databases, a better recognition rate is obtained for MLP classifier. Since the best results were obtained using MLP classifier and the number of speakers and hence the number of speech samples is more in isolated words database with 1000 speakers and 20 words, the results obtained from this experiment is selected. The speech recognition experiment using MLP is repeated by changing a number of parameters by trial and error experiment. The learning parameters used for both LPC and MFCC which produced the maximum recognition rate are given in table 3.9.

Table 3.9: Learning parameters used for LPC and MFCC using MLP

Parameters	Values	
	LPC	MFCC
Number of hidden layers	1	1
Number of neurons in the hidden layer	5	6
Learning rate	.01	.01
Momentum	0.3	0.3
Activation function	Sigmoid function	Sigmoid function

The comparison of the results obtained using LPC and MFCC is evaluated using three measures namely Accuracy, Precision and Recall. Precision and Recall are two other measures which are used to measure the performance of a speech recognition system.

Precision - Precision is the proportion of the examples which truly have class *X* total classified as class *X*. Precision can be expressed using the following equation.

$$precision = \frac{number\ of\ true\ positives}{number\ of\ true\ positives\ +\ false\ positives} \tag{3.29}$$

Recall - It is similar to false positive where proportion of examples which were classified as class *X*, among all examples which truly have class *X*.

$$recall = \frac{number\ of\ true\ positives}{number\ of\ true\ positives\ +\ false\ negatives} \tag{3.30}$$

The table 3.10 shows the performance of both LPC and MFCC in recognising spoken isolated words in Malayalam using MLP classifier.

Table 3.10: Performance analysis of LPC and MFCC

Feature extraction method	Accuracy	Precision	Recall
LPC	79.89	0.811	0.799
MFCC	81.77	0.821	0.818

The comparison of results obtained using LPC and MFCC is given in Figure 3.13.

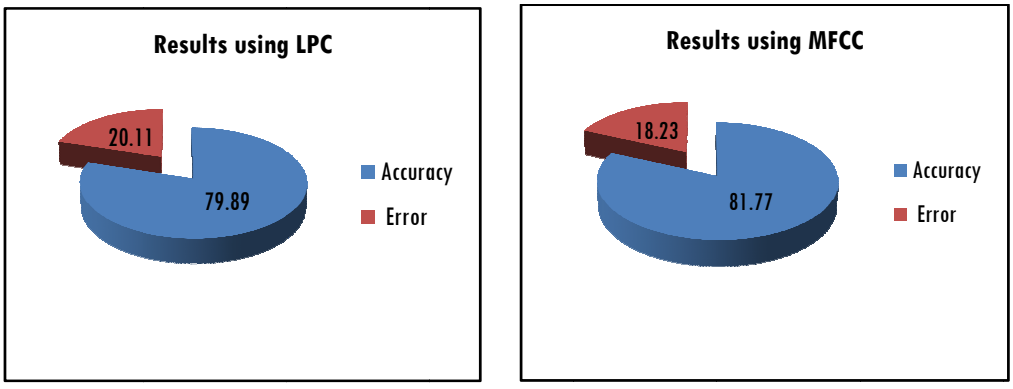


Figure 3.13: Comparison of results obtained using LPC and MFCC

3.7 Summary of the Chapter

This chapter describes the speech recognition systems developed using spectral analysis techniques like LPC and MFCC. Classification algorithms like ANN, SVM, Naive Bayes and HMM were considered. The schemes developed were tested on the Malayalam databases created. The experiments conducted show that both methods are good in recognising speech. In addition to this, three more inferences were obtained from the eight experiments performed which are as follows:

a) *Performance of the speech recognition system degrades when the number of speakers is more.*

b) *Between LPC and MFCC features, MFCC features produced better results on all the three databases.*

c) *MLP architecture of ANN outperformed other classifiers.*

Though the speech recognition systems developed using LPC and MFCC produced a reasonable recognition rate, there is still room for further improvement. So the next chapter discusses the design of speech recognition system using two wavelet based techniques namely DWT and WPD.

Chapter 4

SPEECH RECOGNITION USING WAVELET BASED FEATURE EXTRACTION TECHNIQUES

Contents

In the previous chapter, speech recognition systems developed by using the spectral analysis techniques namely LPC and MFCC with four different pattern classifiers were discussed. This chapter discusses the speech recognition process using two wavelet based feature-extraction techniques namely Discrete Wavelet Transforms (DWT) and Wavelet Packet Decomposition (WPD). This chapter also includes a brief review of wavelets, different wavelet families and wavelet denoising techniques used in this work. Here, nine experiments are carried out. One experiment for selecting the optimal wavelets and eight experiments for developing speech recognition systems using DWT and WPD along with four classifiers.

4.1 Introduction

Wavelet theory is a relatively new development which has become a versatile tool for signal analysis, image processing, speech processing as well as compression. Before the invention of wavelets, Fourier Transforms (FT) and Short Time Fourier Transforms (STFT) were the dominant spectral

analysis tools for frequency domain analysis. The main problem with FT was that FT provided information regarding only amplitude-frequency representation and did not provide any information about the element of time. In order to overcome this limitation, the STFT, which can be considered a Windowed Fourier Transform (WFT), was introduced which was used to provide frequency-time spectrum, by providing time information. But the main limitation of STFT was that it used fixed window size by which the accuracy relied on the size and shape of the window. This greatly affected both frequency and time resolution [114]. So wavelet transforms (WT) were developed in order to overcome the shortcomings of FT and STFT related to its time and frequency resolution problems and it provided a concise and easier analysis of speech signals which is suitable for the non stationary nature of the speech signal.

A wavelet is a versatile mathematical tool that decomposes the signal into its different frequency components where each frequency component is represented with a resolution that matches with its scale called Multi Resolution Analysis (MRA) [115]. Wavelets can be defined as translates and dilates of a fixed function [114, 116]. Unlike other waves like sine waves which are symmetric and regular, wavelets are asymmetric and irregular in nature and have zero average value [117, 118]. Wavelets are well suited for speech processing due to their characteristics such as:

a) A*bility to represent a signal in time and frequency simultaneously*
b) *Fundamental vanishing moment property*
c) *Capability to use windows of varying sizes*
d) *Multi-resolutional, multi-scale analysis*
e) *Flexibility and existence of several bases and*

f) Ability to use long time intervals for low frequency information and short intervals for high frequencies [114, 119, 120, 121].

The outline of the chapter is as follows. A brief description of the different wavelet families are presented in section 4.2. Section 4.3 and section 4.4 presents the design of speech recognition systems using DWT and WPD respectively. The experiments done for finding the optimal wavelets and implementation procedures for DWT and WPD are explained in section 4.5. Section 4.6 analyses the results obtained and the subsequent section provides a comparison of both the techniques. The chapter concludes with section 4.8.

4.2 Wavelet Families

The main idea of wavelet analysis is developed on the theory that every vector in a vector space can be written as a linear combination of the basis vectors in that vector space. Wavelets are functions generated from one single function or basis function called the *prototype* or *mother wavelet* by *dilations* (*scaling*) and *translations* (*shifts*) in time (frequency) domain [114, 122]. If the mother wavelet is denoted by $\psi(t)$, the other wavelets $\psi_{a,b}(t)$ can be represented as

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right) \tag{4.1}$$

Where a is the scaling parameter and b is the shifting parameter. Hence, the parameter a causes contraction of $\psi(t)$ in the time axis when $a < 1$ and expansion or stretching when $a > 1$ [114].

This section provides a very short description of the different wavelet families which are employed in this research work. There are different types of wavelet families [123] available for wavelet analysis such as Haar, Daubechies, Biorthogonal, Symlets, Coiflets, Morlet, Mexican Hat and Meyer.

Wavelets differ in the length of support of the mother wavelet, the number of vanishing moments, the symmetry or the regularity as well as the existence of a corresponding scaling function [119]. Since all the translations and scaling are through the mother wavelet, the selection of the mother wavelet plays an important role in obtaining good recognition accuracy in a speech recognition system. A brief review of the wavelet functions that support DWT and WPD which are tested in this work are given below [117].

- **Haar Wavelets:** The Haar wavelets or db1 is the simplest and oldest type of wavelets which have the shortest support among all orthogonal wavelets. It is a sequence of rescaled, square-shaped function transitions.

- **Daubechies Wavelets (db):** These wavelets, that are very popular, are however, not symmetric in nature and use overlapping windows [123]. Many researches in speech recognition are based on this type of wavelet because they represent a collection of orthogonal mother wavelets which are characterised by a maximum number of vanishing moments for some given length of support. Vanishing moments limit the wavelet's ability to represent polynomial behaviour or information in a signal. They are suitable for denoising signals and for compression [124] since the high frequency coefficient spectrum reflects all high frequency changes. Daubechies wavelets db2 - db20 (even index numbers only) are the commonly used wavelets.

- **Coiflets Family**: It has the highest number of vanishing moments. They are compactly supported and orthogonal but are near from symmetry. Coiflets family members ranges from ‘coif1’ to ‘coif5’.

- **Symlets Family**: These are symmetrical wavelets which were invented to modify the symmetry property of the 'db' family. This family is almost symmetrical, orthogonal, compactly supported in time and has *N* vanishing moments. There are seven members in this family staring from 'sym2' to 'sym8'. Figure 4.1 shows the plots of Daubechies 6, Haar, Coiflet 3 and Symlet 6 wavelets [125].

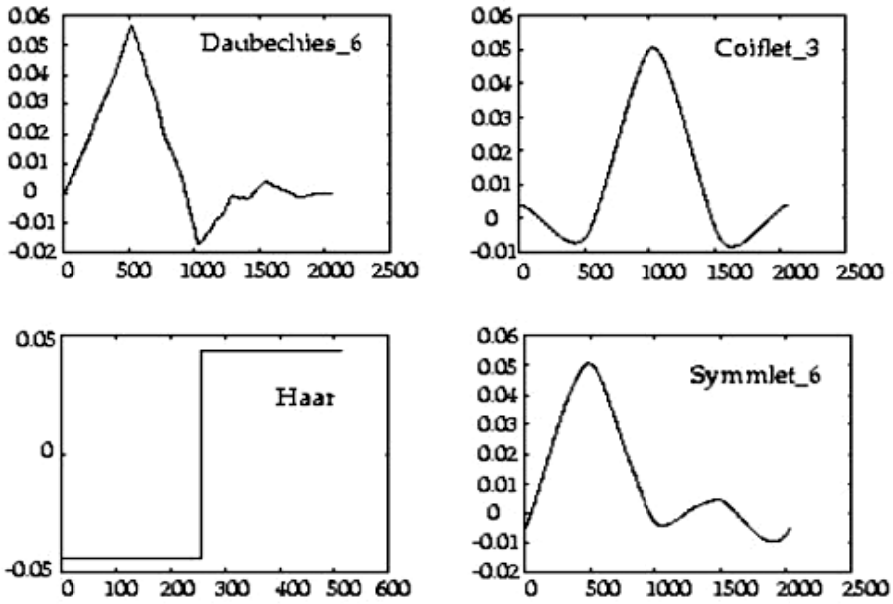


Figure 4.1 Plots of Daubechies 6, Haar, Coiflet 3 and Symlet 6 wavelets

4.3 Speech Recognition System using DWT

DWT provides high time resolution and low frequency resolution for high frequencies and high frequency resolution and low time resolution for low frequencies which are similar to the time-frequency resolution characteristics of human perception system [120, 126]. This section explains the different steps in designing a speech recognition system using the DWT model. As discussed in chapters 2 and 3, here too the recognition scheme is

designed using the four stages of development of the speech recognition system. The different techniques used for the design are given below.

4.3.1 Pre-processing

Pre-processing is a crucial step in the development of a reliable speech recognition system. This includes segregating the voiced region from the silence portion of the captured signal and removing the noise and disturbances from the signal [82]. In this work, two techniques for pre-processing namely *a) End Point Detection and b) Wavelet Denoising* have been used.

4.3.1.1 End Point Detection

This is used to find the start and end point of a signal and to distinguish voiced and unvoiced sounds [81]. The same procedure using ZCR which was applied to LPC and MFCC is exploited here.

4.3.1.2 Wavelet Denoising

The speech signals recorded are often distorted by background noise. So the effect of these noise contents should be removed before extracting the features. There are a number of techniques available for speech enhancement. Here, wavelet denoising techniques are applied to remove the noise contents from the signals. For removing noise from a speech signal using wavelets, three selections are to be performed namely

a) *Selection of wavelet*
b) *Choice of threshold limit and*
c) *The level of decomposition.*

The wavelet threshold denoising algorithm is adopted in this work. There are two popular thresholding functions used for denoising signals using wavelets [127]. They are:

Hard Thresholding (HT): Here, the elements whose absolute values are less than the threshold are set to 0. Hard Thresholding is expressed as

$$X_{Hard} = \begin{cases} X & if \quad |X| > \tau \\ 0 & if \quad |X| \leq \tau \end{cases} \tag{4.2}$$

where X represents the wavelet coefficients and τ is the threshold value.

Soft Thresholding (ST): The elements whose absolute values are lower than the threshold are first set to zero. Then the nonzero coefficients shrink towards 0. This is represented as

$$Y_{Soft} = \begin{cases} sign(x)(|x| - \tau) & if \quad |x| \geq \tau \\ 0 & if \quad |x| < \tau \end{cases} \tag{4.3}$$

In this work, wavelet denoising based on Soft Thresholding is adopted because it is proven that noise can be significantly reduced without reducing the edge sharpness. In soft thresholding, a threshold is estimated as a limit between the wavelet coefficients of the noise and those of the target signal [128]. When compared to hard thresholding, the shrinkage of the wavelet coefficients by ST to zero reduces the effect of singularities and transients [129]. The threshold value chosen cannot be too high or too low. If it is too high, it may remove the contents of the signal and if too low, denoising may not work properly. One of the standard methods for selecting the value of τ is to choose the universal threshold [130] which is defined as

$$\iota = \sigma\sqrt{2\log(N)} \tag{4.4}$$

where σ is the standard deviation of noise and N is the length of the signal which denotes the number of samples in the signal. Standard deviation σ can be calculated as

$$\sigma = MAD / 0.6745 \tag{4.5}$$

where MAD is the median of the absolute value of the wavelet coefficients. Wavelet denoising is considered to be a non-parametric method. For performing wavelet denoising using ST, first an Additive White Gaussian Noise (AWGN) is added to the signal. Suppose *s(t)* is the original signal and the noise added is *n(t),* then a signal *x(t)* can be represented as the summation of the original signal and the noise as [131, 132]

$$x(t) = s(t) + n(t) \tag{4.6}$$

The algorithm for denoising the signals using ST are given in the table 4.1.

Table 4.1: Steps for wavelet de-noising using Soft Thresholding.

1. Choose the mother wavelet from the wavelet family.
2. Select the level N up to which decomposition is to be performed.
3. Calculate the threshold value based on soft thresholding method using equation 4.4
4. For each signal perform the following
   4.1 Compute the wavelet decomposition of the noisy signal.
   4.2 For level from 1 to N, shrink the detail wavelet coefficients by applying soft thresholding technique to reduce the noise using the equation 4.3.
   4.3 Compute inverse wavelet transform to reconstruct the signal based on the original approximation coefficients of level N and the modified detail coefficients of levels from 1 to N

The signals obtained after pre-processing contain the denoised signals which are then applied to feature extraction techniques for extracting the features. The detailed procedure for denoising is explained in chapter 6.

4.3.2 Feature Extraction using DWT

DWT uses digital filtering techniques to obtain a time-scale representation of the signals by decomposing the signal in a more efficient manner than that of traditional bank of filters model and extracts the relevant features [75]. In DWT, the one-dimensional speech signal passes through two discrete-time low and high pass Quadrature Mirror Filters (QMF). The relation between the low pass and high pass filters can be expressed as

$$g[L-1-n] = (-1)^n . h[n] \tag{4.7}$$

where *g[n]* is the high pass filter, *h[n]* is the low pass filter and *L* is the filter length. One filter has a magnitude response which is the mirror image about $\pi/2$ of that of the other filter. This generates the corresponding wavelet coefficients which are represented by two sequences [38]. One sequence represents the low frequency or approximation coefficients denoted by *A* and the second one represents high frequency or detail coefficients represented by *D* [35]. In speech signals, low frequency components have more importance than high frequency signals since the low frequency components represent the characteristics of a signal more than its high frequency components [58, 126]. Figure 4.2 shows the decomposition procedure using DWT.

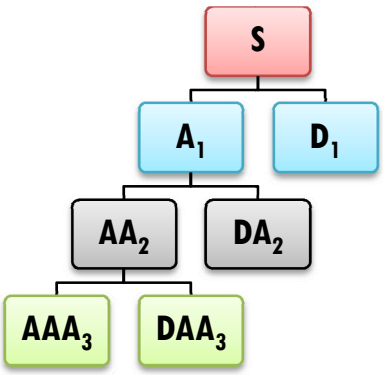


Figure 4.2: Decomposition tree of DWT up to 3 levels

The Discrete Wavelet Transform [125, 133] is defined by the following equation

$$W(j,K) = \sum_{j}\sum_{k} x(k)2^{-j/2}\psi\left(2^{-j}n-k\right) \tag{4.8}$$

where $\Psi(t)$ is the basic analysing function, called the mother wavelet. The functions with different regions of support that are used in the transformation process are derived from the mother wavelet. The original signal *x[n]* is first passed through a half band *high pass* filter *g[n]* and a *low pass* filter *h*[*n*]. This constitutes one level of decomposition and the successive high pass and low pass filtering of the signal can be obtained by the following equations.

$$Y_{high}[k] = \sum_{n} x[n]\cdot g[2k-n] \tag{4.9}$$

$$Y_{low}[k] = \sum_{n} x[n]\cdot h[2k-n] \tag{4.10}$$

where Y_{high} (detail coefficients) and Y_{low} (approximation coefficients) are the outputs of the high pass and low pass filters obtained by sub sampling by 2 [134]. The discrete time domain signal is subjected to successive low pass filtering and high pass filtering to obtain DWT [135]. This algorithm is called the Mallat algorithm [136, 137]. At each decomposition level, the frequency band is reduced to half. With this approach, at high frequencies, the time resolution becomes arbitrarily good while the frequency resolution becomes arbitrarily good at low frequencies. The filtering and decimation process is continued until the desired level is reached. The DWT of the original signal is then obtained by concatenating all the coefficients starting from the last level of decomposition [39].

4.3.3 Post Processing

Normalisation method which was used for post processing the feature vectors obtained using LPC and MFCC is employed here.

4.3.4 Classification

Here also the classification is performed using the 4 classifiers ANN, SVM, Naive Bayes and HMM.

4.4 Speech Recognition System using WPD

WPD is a more detailed method since the signal is passed through more filters. WPD allows time domain information to be incorporated with frequency domain information using multiple window durations. Long windows are used when high resolution information is needed and short windows for extracting low resolution information [63]. It allows simultaneous use of long-time interval for low-frequency information and short-time interval for high-frequency information [138]. Among the four stages of the design procedure, the pre-processing, post processing and classification techniques used are the same as that of DWT. So, only the feature extraction process using WPD is explained below.

4.4.1 Feature Extraction using WPD

Like DWT, the original signal passes through two complementary filters, namely low-pass and high-pass filters, and emerges as two signals called approximation coefficients and detail coefficients [139]. WPD is based on wavelet transform and decomposes a signal with the same widths in all frequency bands [140]. In the next level, both the low frequency sub-bands and high frequency sub-bands are decomposed into lower and higher frequency parts. The decomposition procedure is repeated until the desired

level of decomposition is reached. WPD can also provide a multi-level time-frequency decomposition of signals.

DWT performs a one sided dyadic tree composition of signals. But, WPD allows any dyadic tree structure analysis where the approximation as well as detail coefficients are decomposed iteratively up to a certain level chosen [141]. Thus WPD is a more flexible and detailed method than DWT and it also produces good time and frequency resolutions. The main advantage of WPD is that it provides a more detailed time-scale analysis of the speech signal. The decomposition tree of signals using WPD is given in figure 4.3.

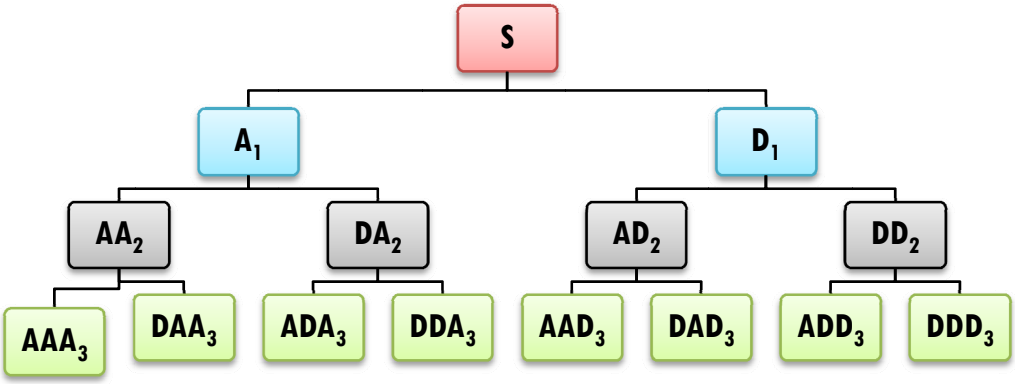


Figure 4.3: Decomposition tree of WPD up to 3 levels

4.5 Experiments

This section presents the different experiments carried out for developing the speech recognition systems using wavelets. Three types of experiments were performed namely:

a) *Selection of optimal wavelets for speech recognition*
b) *Implementation of speech recognition system using DWT and*
c) *Implementation of speech recognition system using WPD.*

4.5.1 Selection of Optimal Wavelets for Speech Recognition

Due to the dominant and powerful properties of wavelets like Multi-Resolution Analysis which analyses the signal at different frequencies giving different resolutions, now-a-days wavelets are extensively used in different fields. There are different types of wavelet families and mother wavelets available. So while using wavelet transforms, an important question which arises is that regarding the choice of the suitable wavelet family; hence the mother wavelet should be used for the analysis of speech signals. Selection of the wavelet family and mother wavelet plays an important role in the performance of the wavelet transforms [119]. So it is necessary to find the optimal wavelet family and the mother wavelet which gives the best recognition accuracy. So experiments are carried out using different wavelet families like Haar, Daubechies, Symlets and Coiflets to locate the best wavelet family and thereby the mother wavelet that is most suitable for the databases created. This is implemented on the database of isolated words with 1000 speakers. The results obtained using different wavelets are given in table 4.2.

Table 4.2: Performance evaluation of a) Haar and Daubechies wavelets, b) Symlets wavelets and c) Coiflets wavelets

Wavelet	Accuracy
Haar or db1	86.9
db2	85.1
db4	87.2
db6	84.5
db8	83.3
db10	82.6
db12	82.0
db14	78.1
db16	81.0
db18	77.8
db20	79.1

Wavelet	Accuracy
sym2	79.5
sym3	80.2
sym4	81.9
sym5	78.3
sym6	78.5
sym7	80.5
sym8	79.3

Wavelet	Accuracy
coif1	75.3
coif2	73.5
coif3	76.2
coif4	76.8
coif5	74.0

The graph given in figure 4.4 shows the performance of different wavelet families on the isolated words database.

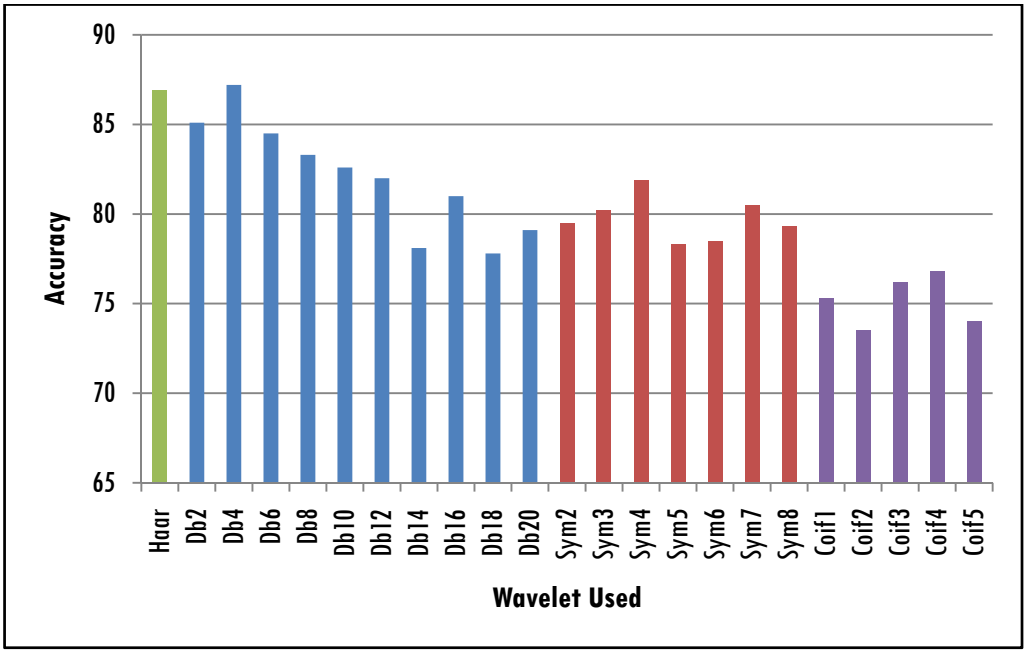


Figure 4.4: Comparison of the performance of different wavelet families

From the experiments conducted using Daubechies wavelets, best results were obtained using db4. Among the symlets wavelets, best results were obtained using sym4 and among the coiflet wavelets, optimal results were obtained using coif4. So it is obvious that Daubechies wavelets which represent the foundations of wavelet based signal processing perform better than the others and the highest recognition rate of 87.2% is obtained using db4 type wavelet. Haar wavelets also performed well on the database with an accuracy of 86.9%.

4.5.2 Implementation using DWT

The algorithm for the implementation of the speech recognition system using DWT is given in table 4.3.

Table 4.3: Algorithm for feature extraction using DWT

1. For each speech signal, find the start and end point of the signal using endpoint detection technique based on ZCR given in eqn. 3.2.
2. Execute the steps in table 4.1 for denoising the speech signals for removing the noise from the signals using Soft Thresholding.
3. Choose the wavelet family and the wavelet (db4) which is appropriate for extracting features.
4. Select the level up to which the decomposition is to be performed.
5. For each speech signal, the following procedures have to be undergone:
   5.1 Decompose the signal into approximation coefficients and detail coefficients.
   5.2 Decompose the approximation signal into new approximation and detail signals.
   5.3 Continue this process iteratively producing a set of approximation signals at different detail levels (scales) and a final gross approximation of the signal up to an appropriate level.

From the experiments performed for finding the optimal wavelets, it was observed that Daubechies wavelets with order 4 outperformed the other wavelets. So, for feature extraction, db4 wavelets are chosen and the decomposition is performed up to 8 levels. Since the recognition rate obtained at the 8^{th} level is count to be better than the other levels, it is chosen as the level of decomposition. For example, the original signal and the eighth level approximation and detail coefficients of two spoken words 'കേരളം' (keralam) and 'താമര' (thamara) are given in figure 4.5 and figure 4.6.

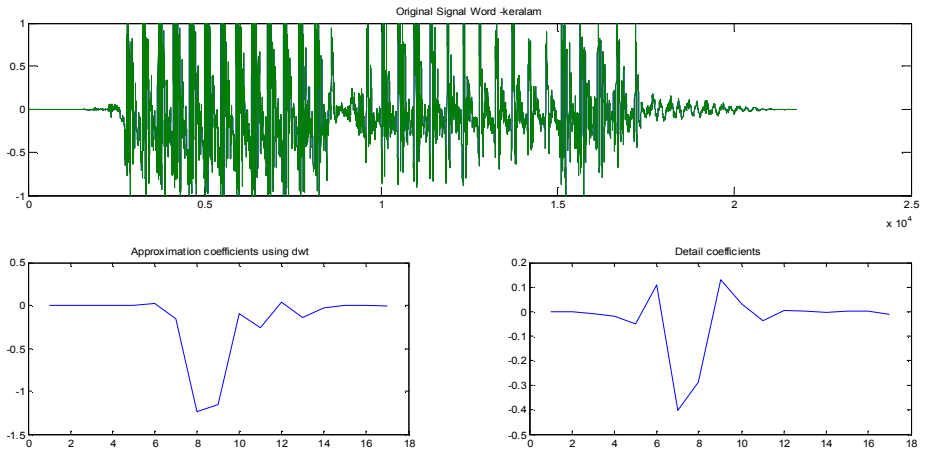


Figure 4.5: Decomposition of word കേരളം (Keralam) using DWT

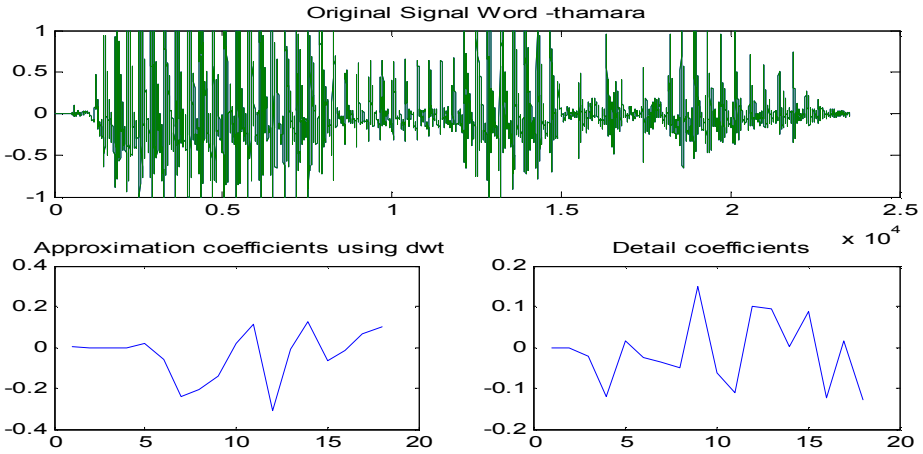


Figure 4.6: Decomposition of word താമര (thamara) using DWT

4.5.3 Implementation using WPD

The first four steps involved in the implementation of the speech recognition system using WPD are similar to that of the algorithm for feature extraction using DWT discussed in table 4.3. So only the wavelet decomposition steps for feature extraction using WPD is explained here which is given in table 4.4.

Table 4.4: Wavelet decomposition steps using WPD

1. For each speech signal, the following procedures have to be undergone:

   1.1 Decompose the signal into approximation coefficients and detail coefficients.

   1.2 Decompose the approximation signal into new approximation and detail signals.

   1.3 Decompose the detail coefficients into new approximation and detail signals.

   1.4 Continue this process iteratively until a specified level is reached.

The experiments using WPD also employs db4 type of wavelets and the decomposition is carried out up to 8th level. The original signal and the eighth level decomposition coefficients of the spoken words 'ചിരി' (chiri) and 'ഗീതം' (geetham) using WPD are given in figure 4.7 and figure 4.8.

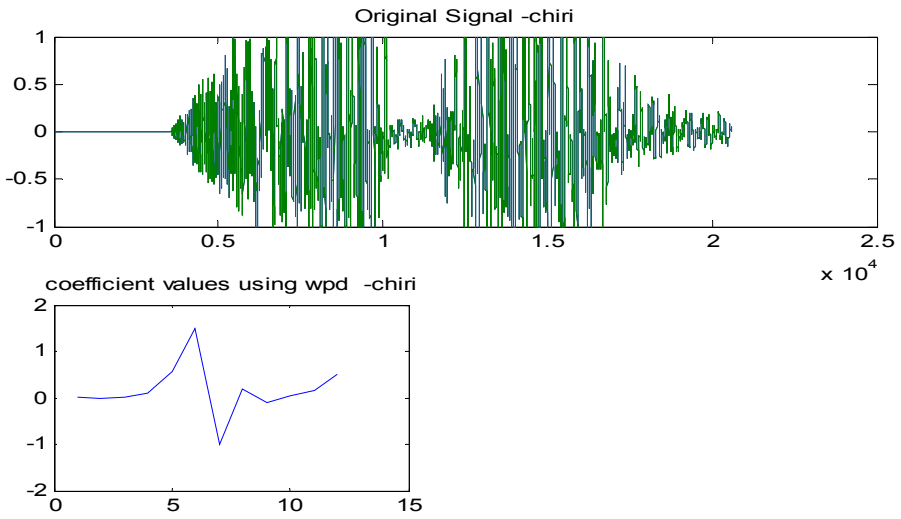


Figure 4.7: Decomposition of word ചിരി (chiri) using WPD

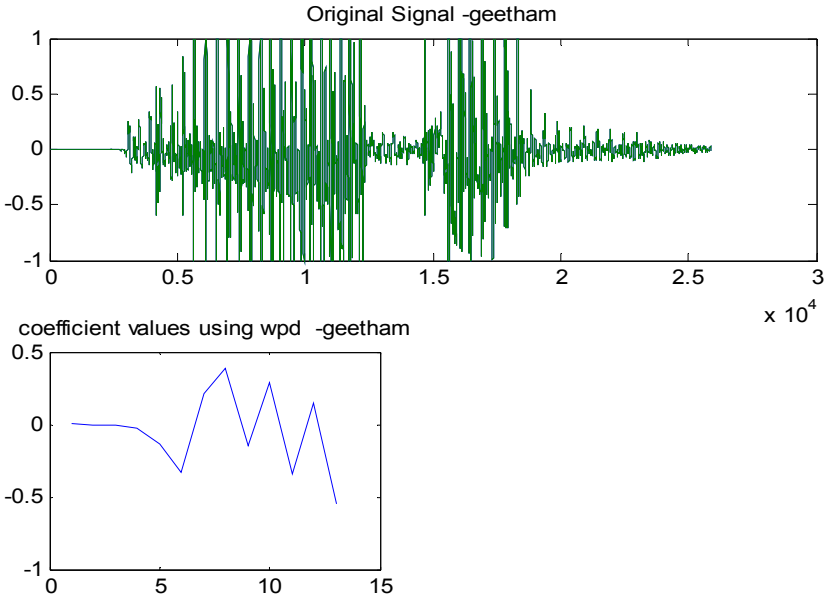


Figure 4.8: Decomposition of word ഗീതം (geetham) using WPD

4.6 Performance Evaluation

The feature vector sets obtained after performing DWT and WPD are normalised and are given for pattern classification using the four classifiers namely ANN, SVM, HMM and Naive Bayes classifiers. This section explains the results obtained for DWT and WPD after classification. The metrics used for performance evaluation are a) Recognition Accuracy, b) Comparison Graphs and c) Confusion Matrix.

4.6.1 Performance Evaluation of Speech Recognition System using DWT

After feature extraction using DWT, 12 features were obtained for each sample. Then these were applied to the pattern classifiers for proper classification into different classes. The experiments were done by changing the number of speakers in order to estimate the impact of variable number of speakers in the recognition rate.

a) Recognition Accuracy

The experiments are done by varying the number of speakers to evaluate the impact of the number of speakers in the recognition rate. The table 4.5, table 4.6 and table 4.7 show the results of DWT obtained after classification using the 4 classifiers.

Table 4.5: Classification results for DWT using MLP, HMM, Naive Bayes and SVM classifiers on Vowels database

No. Of Speakers	Total Samples	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	120	119	99.16	117	97.5	118	98.33	119	99.16
25	300	295	98.33	286	95.33	288	96.0	292	97.33
50	600	583	97.16	564	94.0	568	94.66	577	96.16
75	900	865	96.11	833	92.55	838	93.11	849	94.33
100	1200	1135	94.58	1095	91.25	1108	92.33	1122	93.50

Table 4.6: Classification results for DWT using MLP, HMM, Naive Bayes and SVM classifiers on Digits database

No. Of Speakers	Total Samples	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	100	98	98.0	95	95.0	96	96.0	97	97.0
25	250	243	97.2	236	94.4	238	95.2	241	96.4
50	500	480	96.0	468	93.6	473	94.6	477	95.4
75	750	713	95.06	695	92.66	700	93.33	709	94.53
100	1000	941	94.10	917	91.7	922	92.2	933	93.30
150	1500	1393	92.86	1350	90.0	1368	91.2	1378	91.86
200	2000	1820	91.0	1768	88.4	1783	89.15	1808	90.4

Table 4.7: Classification results for DWT using MLP, HMM, Naive Bayes and SVM classifiers on Isolated Words database

No. Of Speakers	Total Samples	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	200	199	99.5	196	98.0	196	98.0	198	99.0
50	1000	988	98.8	969	96.9	973	97.3	980	98.0
100	2000	1948	97.4	1904	95.2	1907	95.35	1925	96.25
250	5000	4775	95.5	4631	92.62	4671	93.42	4726	94.52
500	10000	9302	93.02	8916	89.16	9031	90.31	9245	92.45
750	15000	13536	90. 24	12998	86.65	13281	88.54	13398	89.32
1000	20000	17442	87.21	16602	83.01	16825	84.12	17217	86.08

b) Comparison Graphs

The graphs given in figure 4.9, figure 4.10 and figure 4.11 show the performance of different classifiers using DWT on the 3 databases.

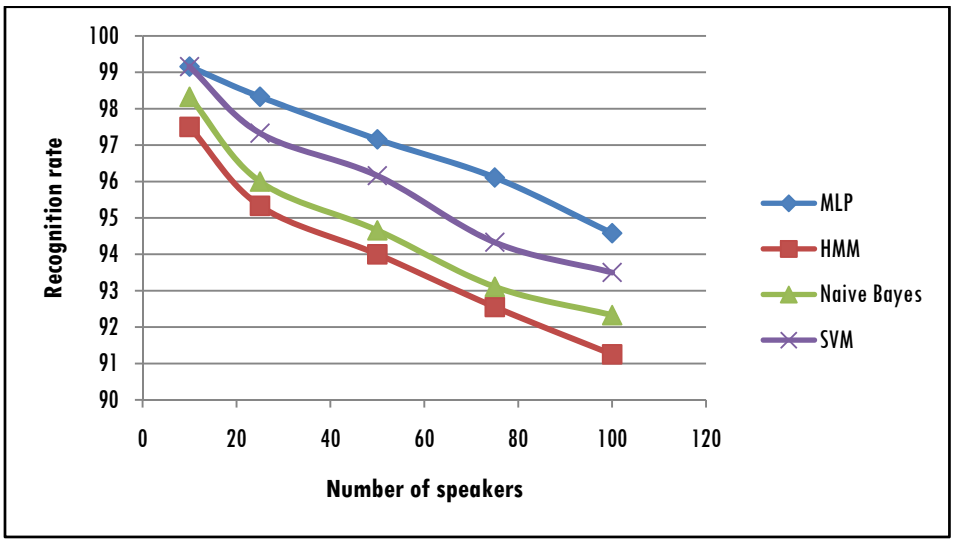


Figure 4.9: Performance of different classifiers on Vowels database using DWT

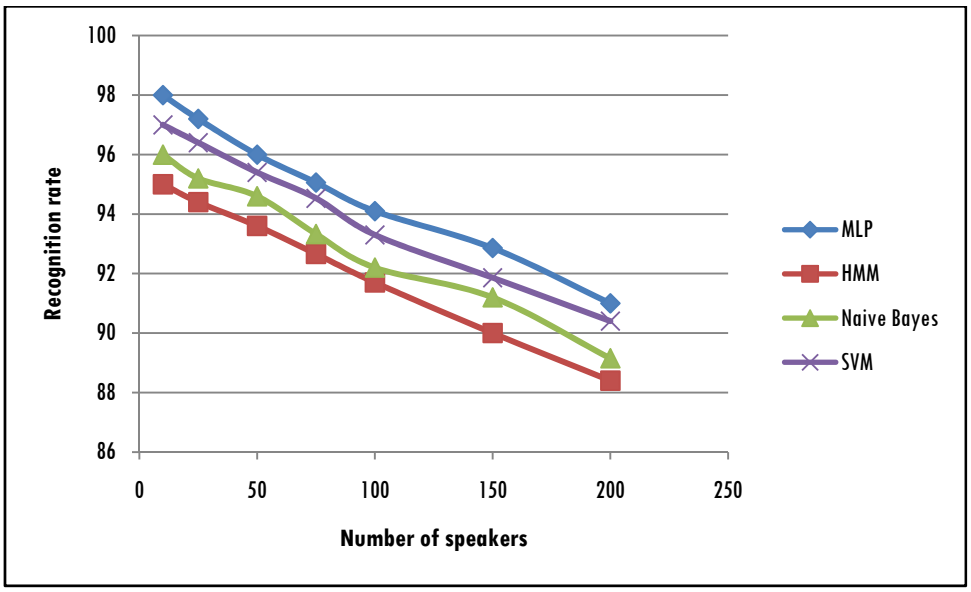


Figure 4.10: Performance of different classifiers on Digits database using DWT

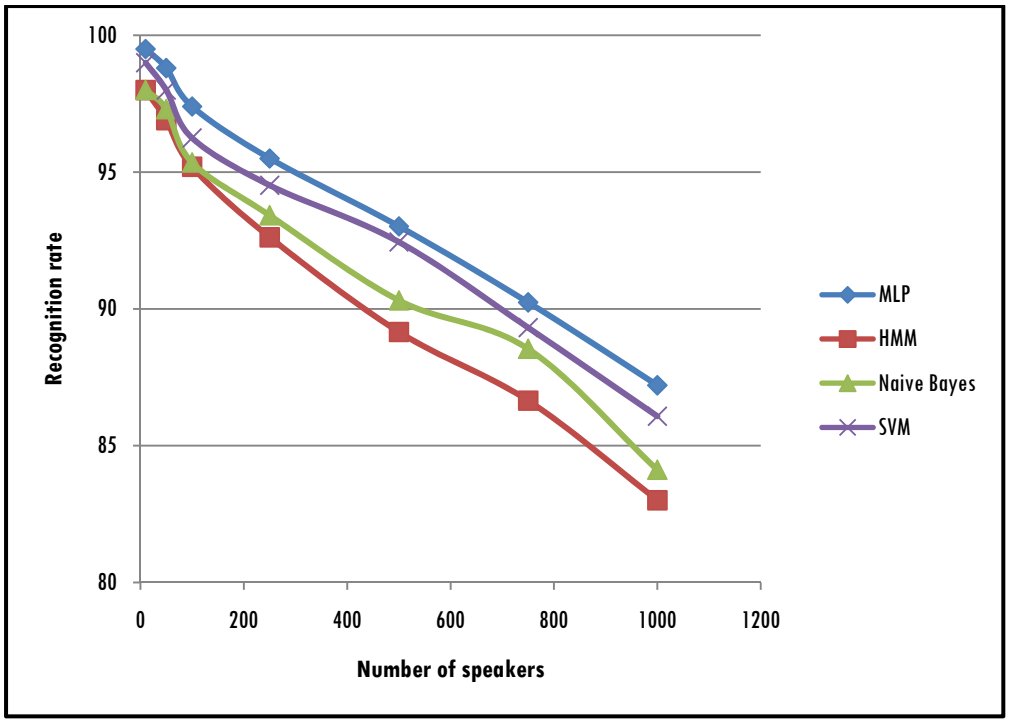


Figure 4.11: Performance of different classifiers on Words database using DWT

From the results obtained using DWT, it is found out that MLP structure of the ANN classifier produces optimal results for all the three databases.

c) Confusion Matrix

Since better results are obtained using the MLP classifier, the results obtained using MLP are selected to represent the confusion matrix. An overall recognition accuracy of 87.21% is obtained for DWT and MLP combination. The figure 4.12 given below shows the confusion matrix generated for isolated spoken words with 1000 people uttering 20 words.

Predicted Class

Actual Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	753	15	20	27	2	23	20	25	4	16	18	6	14	6	21	2	13	3	2	10
2	0	932	4	0	1	0	15	0	6	2	10	0	0	4	3	0	0	5	4	14
3	26	0	895	0	0	0	0	10	0	0	0	10	0	0	0	2	30	22	0	5
4	0	0	0	1000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	4	0	928	0	0	16	0	0	16	0	0	16	0	0	0	5	0	15
6	25	0	0	19	0	863	0	2	0	2	0	0	0	40	2	0	23	4	0	20
7	7	25	0	22	0	10	797	15	0	24	2	18	5	25	0	31	2	1	16	0
8	30	0	19	15	21	0	12	801	1	1	0	8	0	5	25	8	0	14	40	0
9	0	20	25	0	10	0	11	14	696	20	0	1	66	12	0	0	52	71	1	1
10	0	11	0	0	0	3	0	9	0	933	0	0	0	14	4	0	16	0	10	0
11	0	0	4	0	0	15	0	8	0	0	964	0	0	0	0	4	0	0	5	0
12	0	0	0	12	0	0	0	9	0	0	0	959	7	0	0	5	0	0	8	0
13	16	0	0	0	10	0	0	6	0	11	0	0	923	0	14	0	12	0	0	8
14	10	25	23	19	0	15	6	15	0	14	35	8	0	790	0	1	0	24	15	0
15	0	0	10	0	0	14	0	13	0	6	0	15	0	0	905	0	2	7	0	28
16	35	0	16	3	0	4	0	0	10	0	0	17	32	0	22	821	0	8	25	7
17	10	0	0	24	0	50	7	0	6	0	20	0	0	30	24	0	750	64	15	0
18	0	0	12	0	0	32	0	2	0	3	0	6	0	0	0	1	37	887	5	15
19	15	0	32	2	0	3	0	10	0	6	27	2	0	0	17	0	9	4	871	2
20	3	0	0	5	0	0	0	7	0	0	0	3	0	1	1	0	0	6	0	974

Figure 4.12: Confusion matrix for Isolated Words database using DWT MLP combination

4.6.2 Performance Evaluation of Speech Recognition System using WPD

The results obtained using WPD and classification using the 4 classifiers on each database based on recognition accuracy, comparison graphs and confusion matrix are explained in this section. The results obtained after classification based on correctly classified samples and recognition accuracy are given in table 4.8, table 4.9 and table 4.10.

a) Recognition Accuracy

Table 4.8: Classification results for WPD using MLP, HMM, Naive Bayes and SVM classifiers on Vowels database

No. Of Speakers	Total Samples	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	120	118	98.33	114	95.0	115	95.83	116	96.66
25	300	292	97.33	282	94.0	284	94.66	287	95.66
50	600	575	95.83	557	92.83	561	93.5	568	94.66
75	900	853	94.77	821	91.22	828	92.0	836	92.88
100	1200	1122	93.5	1080	90.0	1092	91.0	1105	92.08

Table 4.9: Classification results for WPD using MLP, HMM, Naive Bayes and SVM classifiers on Digits database

No. Of Speakers	Total Samples	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	100	98	98.0	94	94.0	95	95.0	97	97.0
25	250	242	96.8	234	93.6	237	94.8	239	95.6
50	500	475	95.0	462	92.4	467	93.4	471	94.2
75	750	708	94.4	688	91.73	694	92.53	700	93.33
100	1000	935	93.50	906	90.6	912	91.2	926	92.60
150	1500	1384	92.26	1340	89.33	1351	90.06	1369	91.26
200	2000	1809	90.45	1758	87.9	1774	88.7	1796	89.8

Table 4.10: Classification results for WPD using MLP, HMM, Naive Bayes and SVM classifiers on Isolated Words database

No. Of Speakers	Total Samples	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
10	200	198	99.0	194	97.0	194	97.0	196	98.0
50	1000	979	97.9	954	95.4	963	96.3	970	97.0
100	2000	1938	96.9	1882	94.1	1897	94.85	1919	95.95
250	5000	4722	94.44	4560	91.2	4606	92.12	4658	93.16
500	10000	9183	91.83	8800	88.0	8927	89.27	9045	90.45
750	15000	13356	89.04	12805	85.36	12987	86.58	13155	87.7
1000	20000	17368	86.84	16514	82.57	16752	83.76	17107	85.53

b) Comparison Graphs

The graphs given in figure 4.13, figure 4.14 and figure 4.15 show the results obtained after classification based on number of speakers and recognition rate.

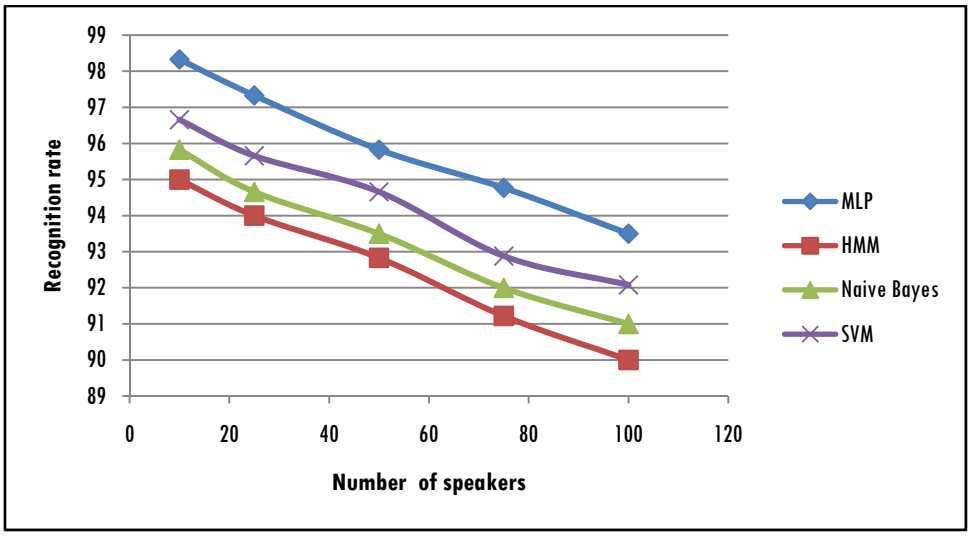


Figure 4.13: Performance of different classifiers on Vowels database using WPD

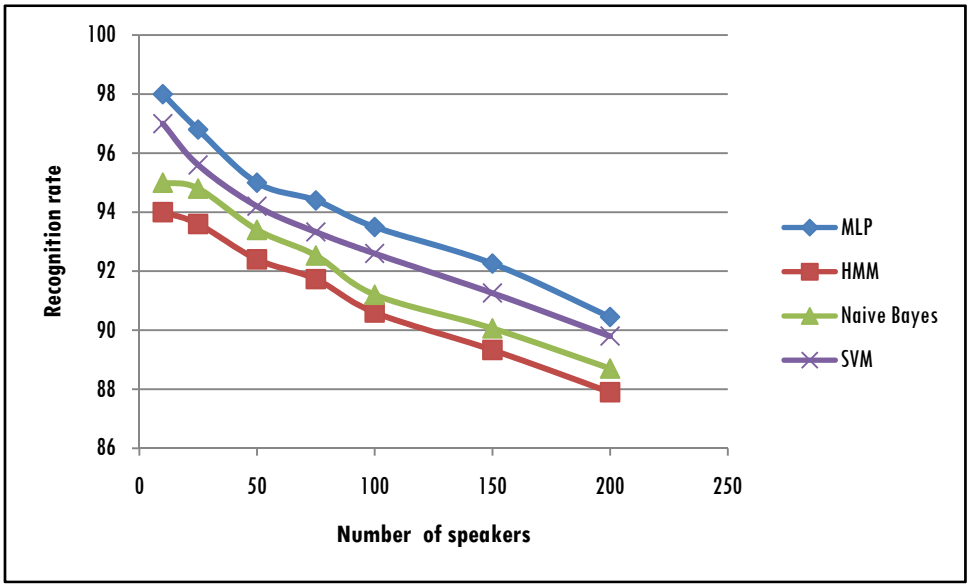


Figure 4.14: Performance of different classifiers on Digits database using WPD

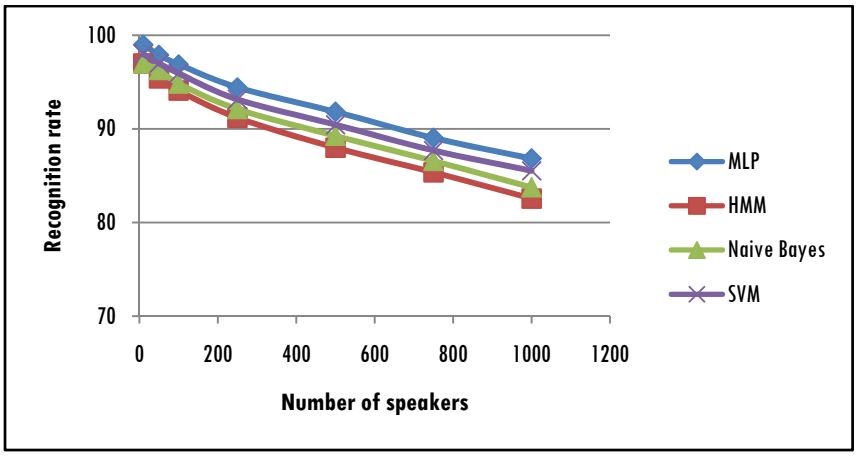


Figure 4.15: Performance of different classifiers on Words database using WPD

c) Confusion Matrix

Predicted Class

Actual Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	927	15	10	2	3	0	4	7	5	1	0	0	2	4	2	2	2	10	2	2
2	3	871	15	7	4	0	0	61	8	0	0	0	0	4	11	3	0	7	0	6
3	0	2	695	0	3	13	1	59	44	5	0	40	0	45	18	1	3	10	60	1
4	2	4	20	927	0	4	2	2	4	0	10	0	5	3	0	2	5	8	0	2
5	0	8	13	4	896	3	3	17	3	2	0	22	2	5	1	0	8	4	0	9
6	0	8	8	0	4	899	2	27	4	0	0	0	0	18	2	1	2	4	0	21
7	20	1	8	32	21	18	795	4	2	2	40	1	1	1	1	2	40	3	7	1
8	0	1	26	0	0	24	2	897	11	1	0	0	2	14	2	2	1	9	0	8
9	22	12	6	31	6	0	13	5	823	0	9	2	5	2	4	17	2	28	8	5
10	0	5	56	0	4	0	2	16	8	866	1	0	2	3	0	24	2	3	0	8
11	1	0	9	0	0	2	0	4	28	9	896	0	0	21	4	11	0	5	0	10
12	0	2	2	17	2	0	0	3	24	0	2	926	3	0	4	10	2	0	1	2
13	0	5	1	2	6	0	0	2	1	0	0	2	968	0	0	7	0	6	0	0
14	0	3	15	0	6	3	0	9	28	16	0	0	2	906	3	2	2	3	2	0
15	18	2	5	60	3	2	40	5	6	40	41	12	1	68	609	54	0	4	26	4
16	2	0	6	17	3	0	2	8	10	17	17	0	0	12	0	882	2	20	0	2
17	0	2	2	0	4	10	4	15	7	2	0	0	4	25	1	2	915	2	0	5
18	0	3	18	3	0	7	33	22	19	0	0	22	0	9	4	2	0	840	0	18
19	2	22	8	0	2	0	0	7	3	0	0	0	4	7	0	25	0	12	903	5
20	0	0	18	0	0	5	0	0	9	0	15	0	0	3	0	10	0	13	0	927

Figure 4.16 Confusion matrix for Isolated Words database using WPD MLP combination

The figure 4.16 given above shows the confusion matrix generated for isolated spoken words with 1000 people uttering 20 words using ANN classifier which generated an overall recognition accuracy of 86.84%.

4.7. Comparison of Performance Evaluation of DWT and WPD

From the results obtained it is found that both the methods are suitable for recognising speech samples. However, it is observed that the results produced using DWT are slightly better than that of WPD in recognising the speech samples from Malayalam. Moreover, WPD decomposes both approximation and detail coefficients. So it is a more detailed method than DWT. So it is computationally more complex than DWT. For all the three databases, better recognition rate is obtained for MLP classifier. The table 4.11 given below shows the performance of both DWT and WPD on isolated words database.

Table 4.11: Comparison of the Performance analysis of DWT and WPD on Words database

Feature Extraction Method	Accuracy	Precision	Recall
DWT	87.21	0.879	0.872
WPD	86.84	0.877	0.868

The comparison of results obtained using DWT and WPD is shown in figure 4.17.

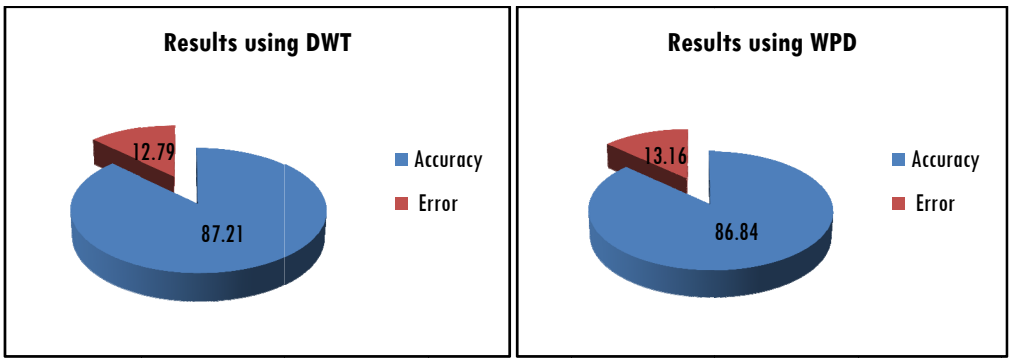


Figure 4.17: Comparison of results using DWT and WPD

4.8 Summary of the Chapter

This chapter discusses the development of speech recognition systems using two wavelet based feature extraction methods namely DWT and WPD using the pattern classifiers ANN, SVM, Naive Bayes and HMM on the three databases created in Malayalam. From the experiments conducted above, 3 conclusions have been arrived at

a) *Among the wavelet based techniques, DWT outperforms WPD*

b) *The MLP structure of ANN is found to produce better results among the classifiers and*

c) *As the number of speakers increases, accuracy is reduced.*

The sixteen different speech recognition systems that were developed for each Malayalam databases using four feature extraction methods namely LPC, MFCC, DWT and WPD and four classifiers namely ANN, SVM, Naive Bayes Classifier and HMM are discussed in chapter 3 and chapter 4. Chapter 5 presents a comparison of the performance of all these speech recognition systems developed and proposes a new enhanced architecture for the better performance of the system.

Chapter 5

COMPARISON OF SPEECH RECOGNITION SYSTEMS AND PROPOSED ENHANCED ARCHITECTURE

Contents

In the previous two chapters, sixteen different speech recognition systems were developed and implemented using a combination of four feature extraction techniques and four classifiers. In this chapter, experimental results obtained using these techniques are thoroughly examined and a comparison is done on the results obtained in order to elicit the feature extraction technique and the classifier combination which produced the best results in terms of recognition accuracy for the databases described in section 2.5. This chapter explains the architecture of an enhanced speech recognition system and also describes the speech recognition system developed using the new feature extraction technique, which can outperform the already tested feature extraction techniques.

5.1 Introduction

A major challenge in developing a novel speech recognition system is to identify the best features that can classify the speech data correctly. So selection of an appropriate feature extraction technique is a crucial factor in

determining the success rate of a speech recognition system. The features selected using different feature extraction techniques are different. Different features produce different results. So the feature vector set obtained plays an important role in determining the recognition accuracy of the ASR system [65]. Another important factor which affects the performance of a speech recognition system is the pattern classifier selected for classifying the feature vectors into appropriate classes.

In chapters 3 and 4, the performance of four feature extraction techniques namely LPC, MFCC, DWT and WPD along with four classifiers such as ANN, SVM, Naive Bayes and HMM have already been evaluated. Though the results obtained are found to be encouraging, there is still room for improvement in the recognition rates obtained. So, improvements and modifications are necessary in the different stages of development of the speech recognition system for an overall better performance. So in this chapter, a new enhanced architecture for designing a speech recognition system with improvements in each stage of development of the system is explained. This chapter also proposes a new enhanced algorithm for feature extraction which is named as *Discrete Wavelet Packet Decomposition (DWPD)* to extract more relevant features and thereby to improve the performance of the speech recognition system.

The rest of the chapter is organised as follows: Section 5.2 provides a comparison of the performance of the different speech recognition systems developed in chapters 3 and 4. Inferences obtained from these sixteen experiments are illustrated in section 5.3. The architecture of the newly designed enhanced speech recognition system is explained in the following section. The architecture of the speech recognition system using the newly proposed feature extraction method is illustrated in Section 5.5. In section 5.6, the implementation procedure

for the newly proposed DWPD algorithm on the isolated words database is explained. Section 5.7 presents the results obtained using this method. A comparison of the results obtained using this proposed feature extraction method with DWT and WPD is performed in section 5.8. The chapter concludes with Section 5.9.

5.2 Comparison and Performance Evaluation of Speech Recognition Systems

In this section the results derived from the speech recognition systems developed in chapter 3 and chapter 4 are compared and evaluated. Since there are three databases, analysis is done on each database separately on the basis of recognition accuracy.

5.2.1 Result Analysis of Vowels Database

The Vowels database consists of 12 vowels and 100 speakers with a total of 1200 speech samples. Since sixteen speech recognition systems were developed using a combination of the four feature extraction techniques namely LPC, MFCC, DWT and WPD and four classifiers namely ANN, SVM, HMM and Naive Bayes classifiers, sixteen different recognition rates were obtained. Table 5.1 shows the performance of these four feature extraction techniques and the four classifiers on the vowels database.

Table 5.1: Comparison of the results of Vowels database based on recognition accuracy

Feature Extraction Technique	No of features	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
LPC	10	1101	91.75	1059	88.25	1072	89.33	1088	90.66
MFCC	12	1109	92.41	1069	89.08	1080	90.0	1100	91.66
DWT	12	1135	94.58	1095	91.25	1108	92.33	1122	93.50
WPD	10	1122	93.5	1080	90.0	1092	91.0	1105	92.08

From the results obtained, it is observed that the recognition accuracy obtained using DWT and MLP combination is better than that of the other combinations with an accuracy of 94.58%. Figure 5.1 shows the graph illustrating the comparison of different feature extraction techniques and classifiers in recognising the vowels stored in the vowels database.

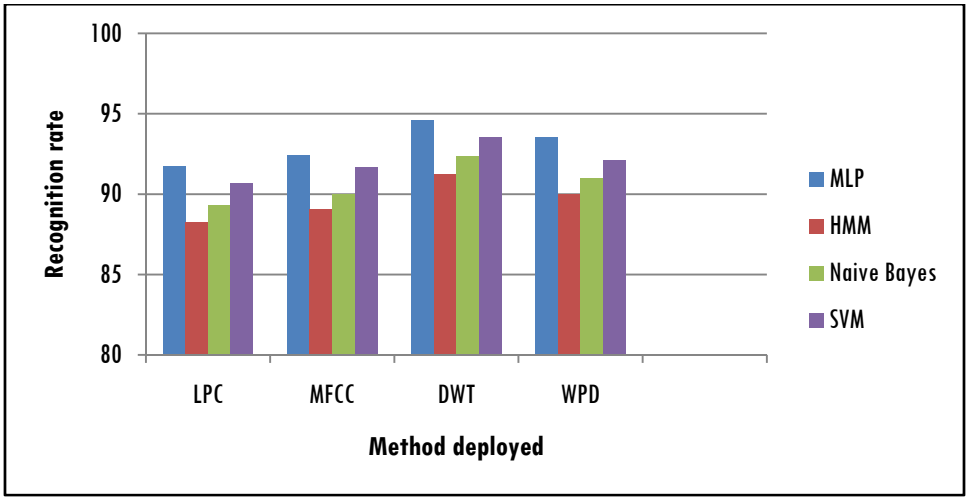


Figure 5.1: Comparison of speech recognition systems on Vowels database

5.2.2 Results Analysis of Digits Database

Digits database consists of 10 digits and 200 speakers with a total of 2000 speech samples. Here also sixteen experiments were carried out. Table 5.2 shows the performance of the above feature extraction techniques and classifiers on the digits database.

Table 5.2: Comparison of the results of Digits database based on recognition accuracy

Feature Extraction Technique	No of features	MLP		HMM		Naive Bayes		SVM	
		Correct	Accuracy	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
LPC	10	1776	88.8	1718	85.9	1734	86.7	1754	87.7
MFCC	12	1794	89.7	1739	86.95	1753	87.65	1776	88.8
DWT	12	1820	91.0	1768	88.4	1783	89.15	1808	90.4
WPD	10	1809	90.45	1758	87.9	1774	88.7	1796	89.8

The results obtained clearly shows that the performance of DWT and MLP combination can recognise the speech patterns more efficiently with an accuracy of 91%. Figure 5.2 shows the comparison graph of different feature extraction techniques and classifiers in recognising the digits stored in the digits database.

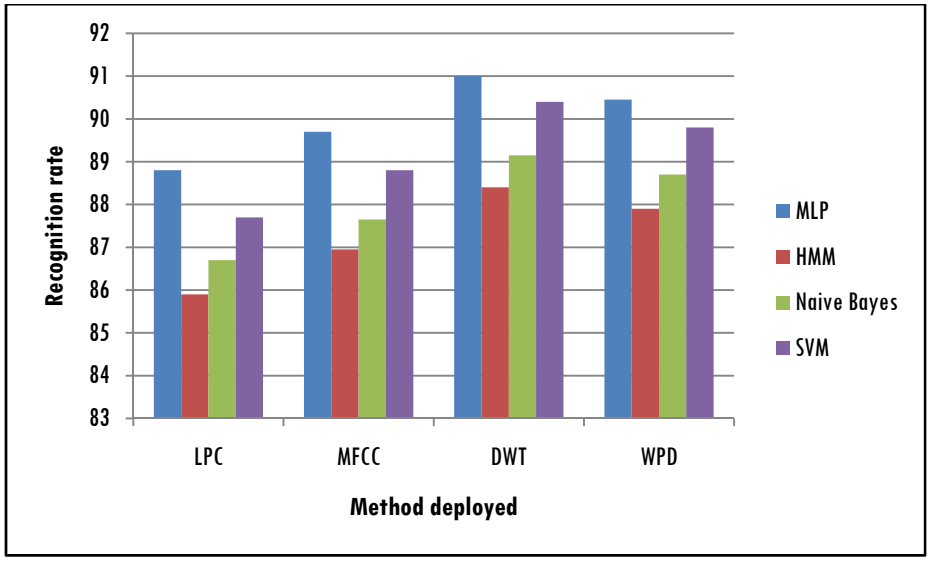


Figure 5.2: Comparison of speech recognition systems on Digits database

5.2.3 Results Analysis of Isolated Words Database

Table 5.3 shows the performance of the four feature extraction techniques and four classifiers on the isolated words database which consists of 1000 speakers and 20 words with a total of 20000 speech samples.

Table 5.3: Comparison of the results of Isolated Words database based on recognition accuracy

Feature Extraction Technique	No of features	MLP		HMM		Naive Bayes		SVM	
		correct	Accuracy	correct	accuracy	Correct	Accuracy	Correct	Accuracy
LPC	10	15978	79.89	15101	75.5	15325	76.62	15602	78.01
MFCC	12	16354	81.77	15452	77.26	15696	78.48	16054	80.27
DWT	12	17442	87.21	16602	83.01	16825	84.12	17217	86.08
WPD	10	17368	86.84	16514	82.57	16752	83.76	17107	85.53

From the sixteen experiments, it is once again seen that DWT and MLP outperforms other combinations with an accuracy of 87.21%. The comparison graph for recognising the 20000 samples of data belonging to isolated words category is given in figure 5.3.

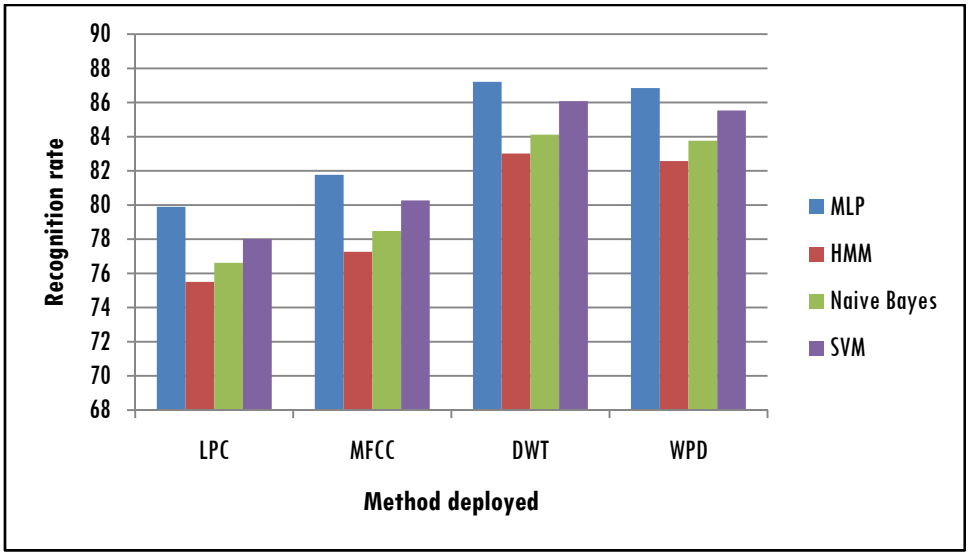


Figure 5.3: Comparison of speech recognition systems on Isolated Words database

5.3 Inferences

In the light of the above results that have been obtained, the following inferences can be generated:

a) Among the spectral and wavelet based feature extraction techniques tested, optimal results were obtained using the wavelet based methods namely DWT and WPD for all the databases.

b) The results obtained using DWT were found to be slightly higher than that of WPD.

c) The MLP structure of the ANN classifier was found to outperform other classifiers.

d) Best recognition rates were obtained using DWT and MLP combination for all the databases.

e) The number of speakers plays an important role in the attainment of recognition accuracy. The recognition rate decreases with increase in the number of speakers.

f) The recognition accuracies obtained from these experiments also reveal the need for an improved system that can effectively handle the large variations present in the speech recognition system.

g) Better generalisations can be obtained when the number of speakers is more, which in turn increases the number of samples in the database.

h) The number of samples in the vowels and digits database are 1200 and 2000 respectively, which is much smaller than the 20000 samples in the isolated words database. From the results obtained, it is obvious that both the vowels and digits databases follow similar results as that of words database. Therefore, further improvements need to be applied only to isolated words database.

5.4 Architecture of the Proposed Enhanced Speech Recognition System

From the above experiments, it was observed that the highest recognition accuracy obtained is for DWT and MLP combination with an accuracy of 87.21% for the isolated words database followed by WPD and MLP combination with an accuracy of 86.84%. So, to further improve the performance of the system, new methods and algorithms are adopted and designed for all the 4 modules of the speech recognition system in the speech recognition process. So changes are made in all the 4 modules of the speech

recognition system for improving the recognition rate thereby improving the performance of the system. The architecture of the proposed enhanced speech recognition system developed is given in figure 5.4.

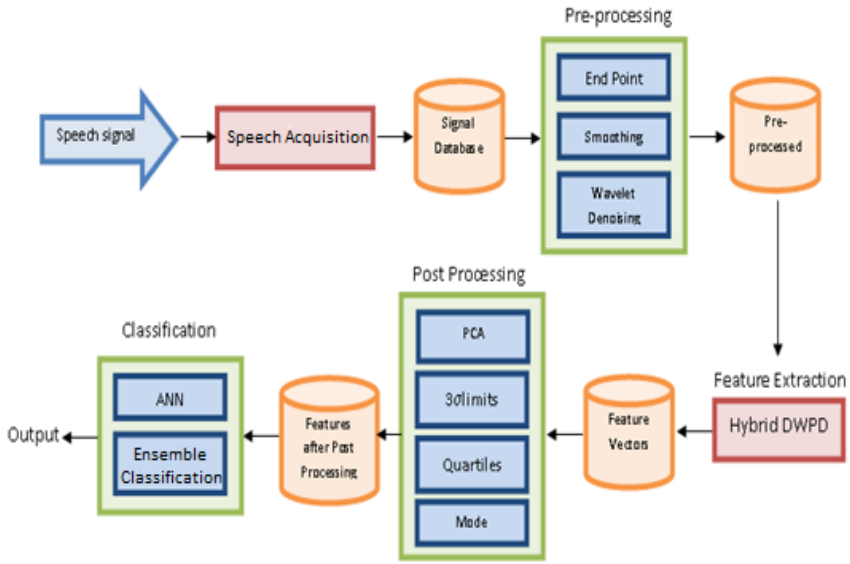


Figure 5.4: Architecture of the proposed enhanced speech recognition system

In the new proposed enhanced architecture, new algorithms and methods are designed, developed and implemented on all the four stages of development of the speech recognition system for improving the overall recognition rate. This chapter presents a new enhanced feature extraction method called Discrete Wavelet Packet Decomposition (DWPD). The following section explains the different modules in the development of the speech recognition system using the proposed DWPD method.

5.5 Speech Recognition using Proposed DWPD Hybrid Method

This section deals with the development of a new feature extraction method for extracting the relevant features from the speech signal. From the

various experiments performed so far, it is found that wavelets are a powerful and extremely useful method for speech recognition. In this section, a new algorithm is proposed for feature extraction which utilises both the features of DWT and WPD which were found to be the best methods so far for the databases. The different steps involved in the development of the speech recognition system using the proposed feature extraction technique for extracting the relevant features is given below which includes the four stages of development of the speech recognition system.

5.5.1 Pre-processing

Since this method is also based on wavelets, the same pre-processing methods which were employed for DWT and WPD are utilised here. The methods used are

a) End Point Detection

b) Wavelet Denoising using Soft Thresholding.

5.5.2 Feature Extraction using DWPD

We have already seen that a wavelet transform decomposes a signal into sub-bands with low and high frequency components. The characteristics of a signal are normally present in the low frequency components, while high frequency components are usually related with noise and disturbance in a signal [142]. Removing the high frequency contents retain the features of the signal and thus reduces the noise in the signal [143]. But sometimes the high frequency components may contain useful features of the signal. The main drawback of DWT is that it cannot decompose the high frequency band any further. Although WPD can achieve this decomposition, it is also applied to

low frequency band signals, which mainly includes the desired signals. This causes unnecessary computational complexity.

To overcome the limitations of DWT and WPD, a new algorithm for extracting the features from the speech signal is designed by combining the features of both DWT and WPD named Discrete Wavelet Packet Decomposition (DWPD). In this method, both DWT and WPD are applied simultaneously. DWT is applied to the low frequency components (approximation coefficients) [114] and WPD is applied to the high frequency components (detail coefficients) [117]. This allows the efficient decomposition of the low frequency components which contain the major signal components using DWT as well as decomposition of the high frequency components using WPD. Since both methods are applied simultaneously, this algorithm utilises the features of both techniques. Figure 5.5 shows the decomposition tree of the proposed DWPD method up to 3 levels.

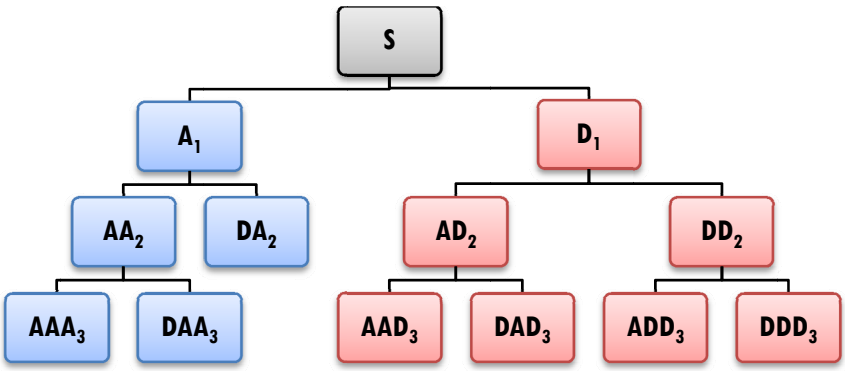


Figure 5.5: DWPD decomposition tree up to 3 levels

The main advantages of the proposed new DWPD hybrid algorithm are:

1. Decomposes high frequency band into more partitions
2. Saves computational complexities
3. Improves recognition rate

5.5.3 Post Processing

The feature vector set obtained after feature extraction using the proposed DWPD algorithm is applied to the post processing techniques for further processing. It is observed that the number of feature vectors obtained is slightly higher than that of DWT and WPD. So the dimension of the feature vector set obtained is reduced using a method called Principal Component Analysis (PCA) [144].

5.5.3.1. Principal Component Analysis

PCA is one of the simplest and most reliable ways of reducing the dimension of data. The main functions of performing PCA are to discover or to reduce the dimensionality of the data set and to identify new meaningful underlying variables. PCA can transform a large set of interrelated data from high to low dimensional space, thus reducing the dimensionality of data without losing the variations in the original data set too much. PCA is applied to extract the most significant components of the feature set obtained. The main objective of PCA is to project the original feature vector onto principal component axis which are orthogonal and correspond to the directions of greatest variance in the original feature space. This reduces redundancy in original feature space and dimensions [60]. The main advantage of PCA is that the extracted features have the minimum correlation along the principal axis [145]. Traditionally, PCA is performed on the symmetric Covariance matrix or on the symmetric Correlation matrix which can be calculated from the feature data matrix.

5.5.4 Classification

Classification is performed using the MLP architecture of ANN since it is established as the best classifier for the databases created.

5.6 Implementation of DWPD Algorithm

The speech signals after pre-processing are applied to the proposed DWPD technique for extracting the features. The outline of the proposed DWPD algorithm is given in table 5.4.

Table 5.4: Algorithm for feature extraction using DWPD algorithm

1. For each speech signal perform the following steps.
1.1 Split the speech signal into two bands by decomposing the signal up to one level to obtain a low frequency band signal and a high frequency band signal.
1.2 Apply 7 scales of DWT on the low frequency components.
1.3 Apply 7 scales of WPD on the high frequency components.
1.4 Combine the features obtained from both these decompositions to form the new feature vector set.

5.7 Experimental Results

After implementing the DWPD algorithm for extracting the relevant features from the speech signals, 18 features are obtained for each sample. Since the number of feature vectors is greater than that of the number of features obtained using DWT and WPD, the feature dimension is reduced using PCA. A speech recognition system is considered to be more efficient and robust if it can recognise the speech samples with a minimum number of features. From the results obtained, it is observed that there is only a slight difference in the recognition rate obtained from classification before and after feature reduction using PCA. The table 5.5 given below shows the comparison of results based on recognition accuracy obtained before feature reduction and after feature reduction.

Table 5.5: Comparison of results before and after feature reduction

	Total samples	No of features	Accuracy %
Before feature reduction	20000	18	89.98
After feature reduction using PCA	20000	14	89.505

Since there is not much difference in the results obtained, the reduced feature vector set is chosen for further classification. Figure 5.6 shows the confusion matrix of the results obtained after performing PCA.

Predicted Class

Actual Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	969	5	0	0	2	0	1	1	0	0	2	4	5	1	0	3	1	1	0	5
2	4	898	0	9	1	0	8	6	2	1	1	10	3	19	2	1	0	3	20	12
3	1	3	975	0	1	2	0	0	0	0	2	2	4	0	1	2	0	1	1	5
4	0	2	0	973	2	3	0	2	1	0	5	0	1	0	4	0	1	0	0	6
5	0	2	0	0	975	0	1	0	1	3	1	0	3	1	3	0	3	0	5	2
6	21	5	12	47	26	704	3	20	28	11	20	4	6	26	2	31	2	11	15	6
7	0	8	1	1	2	0	968	0	0	2	1	1	1	0	2	1	7	2	1	2
8	0	2	0	0	5	0	2	969	5	2	1	1	2	0	1	0	4	0	0	6
9	20	40	15	44	31	45	26	0	575	32	27	4	4	20	18	25	15	14	41	4
10	0	2	0	0	4	0	1	2	1	968	0	2	4	3	2	2	3	2	0	4
11	0	4	0	0	1	1	0	0	0	1	975	2	4	0	2	1	0	1	0	8
12	0	14	0	0	1	2	0	0	0	1	0	971	4	0	2	0	2	0	1	2
13	30	11	17	32	3	11	41	8	23	0	17	2	727	2	11	17	15	5	2	26
14	0	7	0	0	1	0	1	0	0	2	1	1	2	971	2	2	0	0	4	6
15	33	4	22	14	2	15	26	28	10	17	3	16	4	52	652	39	2	35	21	5
16	11	6	13	25	22	1	17	24	16	16	1	33	2	15	23	731	3	2	34	5
17	0	2	0	0	1	1	1	0	3	1	2	0	2	1	1	2	976	0	0	7
18	0	6	0	0	2	1	0	1	0	0	0	0	2	1	3	3	0	975	3	3
19	0	2	0	0	0	0	1	0	0	4	0	2	5	2	3	3	0	0	970	8
20	0	0	0	0	6	0	4	0	0	1	0	4	5	0	1	0	0	0	0	979

Figure 5.6: Confusion matrix for Words database using DWPD MLP combination

5.8 Performance Evaluation

The performance of this new method can be evaluated by comparing the results obtained with that of the results obtained using DWT and WPD alone. Table 5.6 given below shows the performance analysis of these 3 methods.

Table 5.6: Comparison of DWT, WPD and DWPD methods

Feature extraction	Precision	Recall	Correctly classified	Recognition Accuracy %
DWT	0.879	0.872	17442	87.21
WPD	0.877	0.868	17368	86.84
DWPD	0.9	0.895	17901	89.505

The graph in figure 5.7 shows the comparison of results obtained using DWT, WPD and DWPD in terms of recognition accuracy.

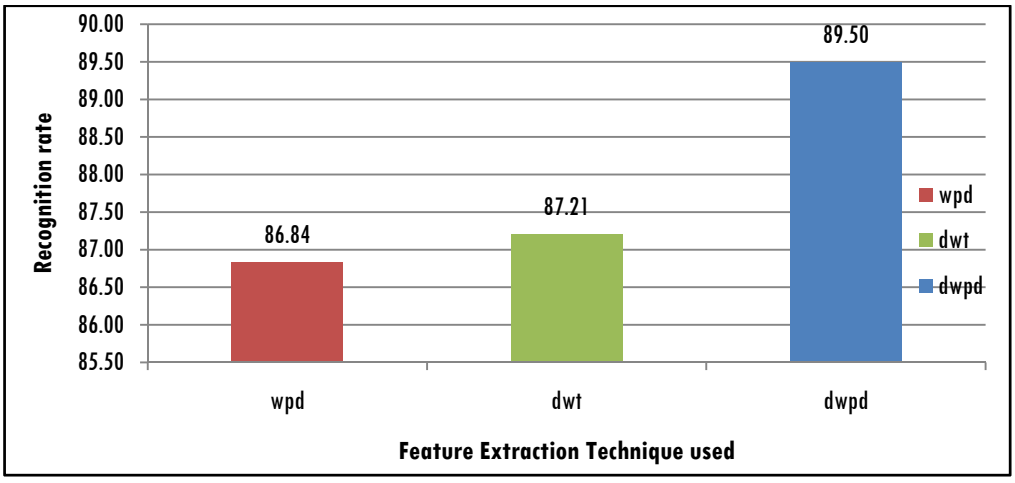


Figure 5.7: Comparison graph of DWT, WPD and DWPD

5.9 Summary of the Chapter

A novel model for deriving feature vector set for speech recognition is introduced in this chapter. The proposed model is based on the wavelet decomposition method which uses a combination of DWT and WPD. The new algorithm is implemented on the words database and the results obtained are

evaluated. The results obtained clearly show that though good recognition accuracy is obtained for DWT and WPD, a hybridised combination of both the techniques can perform better than that of the individual methods when used separately. This chapter also discusses the successful application of the data reduction technique, PCA. In this chapter, the new algorithm DWPD is developed during the feature extraction stage for improving the performance of the speech recognition system. Equally important are the pre-processing, post processing and classification techniques adopted. So the next three chapters will discuss the new techniques and modifications that are designed for these three stages in the development of a speech recognition system.

Chapter 6

SPEECH ENHANCEMENT USING PROPOSED ADAPTIVE SMOOTHING TECHNIQUE

Contents

The implementation of the hybrid algorithm DWPD, discussed in the previous chapter produced improvements in the recognition accuracy. This chapter proposes a new speech enhancement algorithm by smoothing the speech signals during the pre-processing stage for the further enhancement of the speech recognition system. The smoothed signals obtained are then again pre-processed using the wavelet denoising method based on Soft Thresholding and are used for the implementation of the speech recognition system. The hybrid algorithm developed in the previous chapter and the MLP architecture which produced better results than the other classifiers are used for feature extraction and classification respectively.

6.1 Introduction

The quality of the input speech plays an important role in the performance of a speech recognition system. From an engineering perspective,

noise can be considered as a random or pseudo-random signal which is distinguished by how energy is distributed in the spectrum of the signal [146]. Noise may contain uniform energy distribution (white noise) or non uniform energy distribution (coloured noise). Speech signals are affected by either additive or environmental noises like background noise, impulse noise, speaker interfering noise etc. and non additive noises like speaker stress, non-linearities of microphones etc. Recovering data from noise is an important area of research since it affects the recognition accuracy of a speech recognition system. Though many parameters affect the accuracy of the speech recognition system, the presence of speech noise is one of the key factors [147].

Since the main objective of speech is to facilitate easy and accurate information-exchange, the quality and intelligibility of speech is of utmost importance [2]. Degradation of speech due to the presence of different types of noise is a serious issue in speech recognition, speaker recognition and speaker verification. This study is important as the performance of ASR systems degrades dramatically in adverse environments. This greatly limits the speech recognition application deployment in realistic environments. So, this chapter addresses this problem of building a speech recognition system with minimum amount of noise for the better performance of the system. In this chapter, a new algorithm is developed to smooth the signals before applying speech denoising techniques. A hybrid architecture is developed using a combination of the proposed adaptive smoothing algorithm and wavelet denoising method based on ST called Adaptive Smoothing Soft Thresholding (ASST) for removing as much noise from the signals.

The chapter is organised as follows. Section 6.2 explains the concept of speech enhancement. The theory of smoothing is described in section 6.3 and the proposed adaptive smoothing algorithm is presented in the next section.

The procedure for the implementation of the Adaptive Smoothing Soft Thresholding method is explained in section 6.5 followed by the simulation experiments and results obtained using the ASST algorithm in the subsequent section. Section 6.7 illustrates the results obtained after classification using MLP. A comparison of the performance evaluation of this algorithm with Soft Thresholding is described in section 6.8 and section 6.9 concludes the chapter.

6.2 Speech Enhancement

One of the main factors that interfere with the accurate recognition of speech is the presence of background noise. If there is too much of noise from the background, the words cannot be heard or understood properly. A main criterion for measuring the noise present in a signal is by calculating the Signal to Noise Ratio (SNR). A speech signal which is affected by background noise has a low SNR. If the SNR value of a signal is high, it is considered to be noise free. Speech enhancement and additive background noise suppression are of great importance because reducing the amount of noise has utmost importance in developing an efficient speech recognition system [148]. For this, speech enhancement methods and algorithms are used to improve the quality and clarity of the speech by reducing noise [149].

The main objective of speech enhancement is to recover the original signal from noisy data retaining the important properties of the original signal [150]. Now-a-days, signal denoising have become an intensive field of study because of the increasing applications of speech enhancement in the areas like digital mobile radio telephony systems, pay phones in noisy environments, teleconferencing systems, hearing aids, etc. Though the interest in practical and powerful speech enhancement algorithms has grown considerably, and significant progress has been made, speech processing under adverse

conditions is still a challenging area of research [151]. Speech enhancement algorithms are used to improve the performance of communication systems by improving the perceptual quality and intelligibility of speech when they are affected by noise. Restoring the desired speech signals from noise is still an elusive goal in speech communication.

Degradation of signals by noise is a key problem in signal processing. A signal is assumed to vary smoothly most of the time. It is assumed to have only few abrupt shifts. But in real world situations this is not the case. The speech signals are often affected by additive noise. This is a severe problem that greatly degrades the quality, clarity and intelligibility of the speech signals. In order to solve this, it is necessary to enhance the corrupted signal through an appropriate speech enhancement algorithm. Researchers have developed different methods over the years to solve this problem. Traditional denoising schemes are based on linear methods, where the most common choice is the Wiener filtering [152, 153]. Recently, nonlinear methods, especially those based on wavelets have become increasingly popular [154]. Wavelet denoising techniques usually use thresholding techniques like soft thresholding [126] and hard thresholding [128]. Different other thresholding techniques were also developed by selecting new threshold values and by combining different thresholding techniques [149, 150].

Though different variations of threshold values are available, the most popular thresholding techniques that are widely used are Soft Thresholding and Hard Thresholding. We had already discussed these thresholding techniques in chapter 4. Soft Thresholding was used to remove noise from the signals before feature extraction using DWT and WPD in chapter 4. As discussed in chapter 4, in Hard Thresholding, the elements whose absolute values are lower than the selected threshold are set to zero and in Soft Thresholding, the elements whose absolute values are lower than the selected threshold are first set to zero and then

the non zero coefficients are shrinked towards zero. Figure 6.1 shows the wavelet denoising using Hard Thresholding and Soft Thresholding.

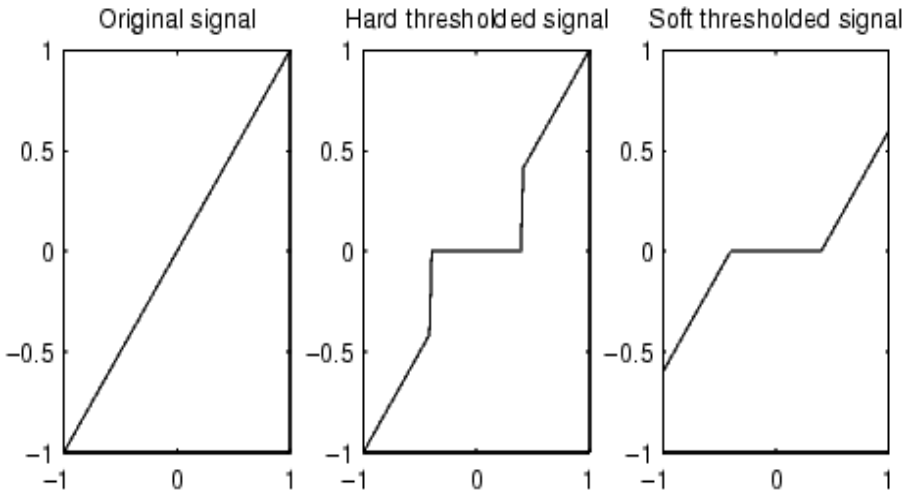


Figure 6.1: Original signal and the signals after Hard and Soft Thresholding

Research done in the area of speech enhancement using Soft Thresholding and Hard Thresholding shows that Soft thresholding produces more Signal to Noise Ratio and less error when compared to Hard Thresholding and is found to be better than hard thresholding in removing the noise from the signals [129]. This is due to the fact that hard thresholding procedure creates discontinuities at the threshold value (when $x = \pm$ threshold value). The soft thresholding procedure takes care of this issue and does not create discontinuities. Though thresholding techniques are found to be good in removing noise, they are not much capable of extracting the real trends in the signals. Hard and Soft Thresholding take into consideration only the value of the signal at that point in time. It does not take into consideration the previous or future values losing out on the trends and directions of the signal thereby accommodating noises and other disturbances which should have not been considered. Suppression of unwanted signals would have been more successful had there been a view on the trend.

6.3 Smoothing of Speech Signals

The basic intention behind the elimination of noise is to enhance the speech signal [155]. Though several speech enhancement algorithms based on thresholding techniques are available, a simple threshold can suppress the noise only up to a limited extent. In many cases the true signal amplitudes which are plotted along Y-axis of a waveform change rather smoothly as a function of the X-axis values. But, in real world situations, many kinds of noise are seen as rapid with random changes in amplitude from point to point within the signal. In such cases, noise can be further reduced by a method called smoothing. In traditional smoothing, the data points of a signal are modified so that individual points that are higher than the immediately adjacent points due to noise are reduced, and points that are lower than the adjacent points are increased. This naturally leads to a smoother signal.

Smoothing techniques are used to eliminate noise and extract real trends and patterns in a speech signal. Smoothing reduces the effect due to random variations in the signal. Rapid fluctuations or volatility in the speech signals hide or mask some information that lies beneath. So these fluctuations are to be reduced by smoothing the signals. Smoothing usually removes high frequencies and retains low frequencies whereas denoising attempts to remove whatever noise is present in the signal and retains whatever signal is present regardless of the frequency content of the signal [156].

Speech signals are often contaminated by sudden, abrupt noise signals that are represented in the form of spikes or troughs. Spikes are abrupt discontinuities in a speech waveform which are caused by the rapid fluctuations or volatility of data in the signal. The spikes/troughs distort the calculations and produce erroneous results. Elimination of such sudden variations has always been a challenging task. Spikes often mask the details of the true data present in the

signal. This may lead to misidentification of the signal. This may subsequently cause error in the quantification of the signal. So in order to avoid these misclassifications, the spike/troughs in the signal should be removed without causing distortion in the rest of the curve. Though there exist several noise-removal algorithms, they are not that efficient for removing spikes. The random errors result in noise and these errors may be due to different reasons. So, different strategies are followed to remove noise from the data [77].

So in this work, a new algorithm is proposed which is used to smooth the signals before applying wavelet denoising techniques. Smoothing is done to eliminate speech like noise components of the speech signal. Smoothing can be non-linear [157] as well as linear. In linear smoothing, the speech signal is passed through a hamming filter with an impulse response and are separated based on their non-overlapping frequency content. Non-linear smoothing separates signals based on whether they are considered as smooth or rough. A popular non-linear smoothing technique is Median smoothing in which L samples of speech is taken at every time instant. The smoothed signal is the median of those L samples. The magnitude of the sudden spikes in the signal is reduced, there by obtaining a signal which is smoother than the original signal. Most of the linear filters confuse and remove the high frequency components of the signals along with noise. So in many cases where there are jumps in the signals, it may cause edge blurring [158]. The main idea of the median filter is to run through the signal entry by entry, replacing each entry with the median of neighbouring entries. Denoising using wavelets are quite different from traditional filtering techniques because it is nonlinear.

6.4 Proposed Adaptive Smoothing Algorithm

Many research activities have been carried out and proposed for finding new speech denoising algorithms [159] using different threshold values [160]

and by combining different thresholding techniques [155]. In this work, we have used the already well defined soft thresholding (ST) technique. The new idea developed is to smooth the signal before applying thresholding. In ordinary thresholding, only the top/bottom portions of a spike are cut out. But the steep gradient in the waveform will exist which actually accentuates the distortion. Attenuating a distortion without actually affecting the original waveform is of prime importance.

In the proposed method, previous values are compared with future values to determine the general trend of the signal and thereby facilitating suppression of random troughs. If these sudden spikes are reduced by smoothing, then automatically more noise components can be reduced and the original signal can be captured in its fullness. The signal after smoothing is then applied to denoising using soft thresholding which produces SNR values which are greater than that of using soft thresholding alone.

This new proposed method can be considered similar to ZCR because it also deals with the sign of the signal. Here, the sign of the present value of the sample Y_i and the next value Y_{i+1} are compared. If both the values are in the same direction and in an increasing trend, the samples are reproduced in total and amplified by a smoothing factor less than 1, say 0.5 which decrease when the trend continues. When there is a reversal in trend, the factor that is added is kept high to capture the reversal in total. If Y_i and Y_{i+1} are in opposite directions, or in other words if there is a sign change in magnitude, we apply a dominant factor limiting the fall. If the trend continues, the signal is again reproduced in total, plus the factor. Table 6.1 explains the algorithm for smoothing the signals using adaptive smoothing technique.

Table 6.1: Steps in Adaptive smoothing Algorithm

1. Initialize smoothing factor as < 1, say 0.5
2. Calculate the length of original signal, l
3. Identify the sign of first point of the signal, prev_sign.
4. For each point in the signal repeat steps for I = 1 to l

4.1 Identify the sign of signal at the current point, say curr_sign.

4.2 Identify the sign of signal at the consecutive point, say next_sign.

*4.3 Compute sign=curr_sign * next_sign*

4.4 If (Sign is positive && curr_sign is negative && next_sign is negative)

Sign = negative

endif

4.5 If (sign is positive && prev_sign is positive)

*Smoothingfactor =smoothingfactor *0.5*

Else

If (sign is positive & previous sign is negative) Smoothing factor = 0.5

else

if (sign is negative && prev_sign is positive)

smoothingfactor = 0.5

else

*Smoothingfactor =smoothingfactor *0.5*

*4.6 Calculate Original-signal (i)=curr_signal + ((next_signal- curr_signal) *smoothingfactor)*

4.7 Assign to prev_sign=curr_sign

5. *The End (of algorithm.)*

Adaptive smoothing [161] technique takes care of the deficiencies in the HT and ST methods. Adaptive smoothing holds the present and future values.

When the trend remains the same, the signal is reproduced as such, removing the minor deviations. When there is a sudden reversal, the trend is checked and in case the reversal trend continues the signal is reproduced smoothly and in case the trend is not continued the sudden dip is smoothed out. To achieve the smoothing, the signal is multiplied by a factor. In case of a temporary change in signal, the factor suppresses the fall to a large extent and smoothens the same. In case the trend continues the signal is continued as such.

Consider the following segment of a signal. When the smoothing algorithm is implemented on speech segments, the sharp edges are smoothed and the resulting signal appears smoother than the original signal. Figure 6.2 shows an example of the original speech segment and its corresponding signal obtained after adaptive smoothing.

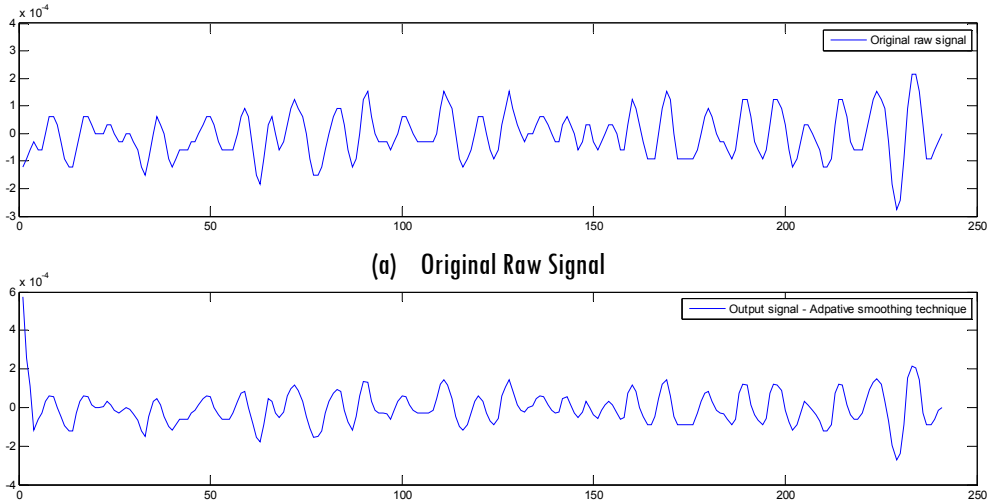


Figure 6.2: (a) Original Raw Signal, (b) Signal after smoothing

Adaptive Smoothing technique which is developed in this work takes care of the deficiencies in the hard and soft thresholding methods. Adaptive Smoothing technique, holds the present and future values. When the trend

remains the same, the signal is reproduced as such, removing the minor deviations. When there is a sudden reversal, the trend is checked and in case the reversal trend continues signal is reproduced smoothly and in case trend is not continued the sudden dip is smoothed out. To achieve the smoothing, the signal is multiplied by a factor. In case of a temporary change in signal, the factor suppresses the fall to a large extent and smoothens the same. In case the trend continues the signal is continued as such.

6.5 Implementation of Adaptive Smoothing Soft Thresholding (ASST)

After performing smoothing, these signals are applied to wavelet denoising using soft thresholding. Since adaptive smoothing is followed by soft thresholding, this method is named as Adaptive Smoothing Soft Thresholding (ASST). The soft thresholding procedure is already discussed in chapter 4. This can be mathematically represented as

$$x(t) = s(t) + n(t) \tag{6.1}$$

Where *x(t)* is the original signal and *n(t)* is the additive noise which are expressed as functions of time *t* [162]. Let *W(.)* and $W^{-1}(.)$ represent the forward and inverse wavelet transform and $W(., \tau)$ denote the wavelet denoising operator with soft threshold value τ where the threshold value is calculated using the equation 4.4 in chapter 4. Then the following three steps can be used to generate $\hat{S}(t)$ from *S(t)* by denoising *X(t)*.

$$Y = W(X) \tag{6.2}$$

$$Z = D(Y, \tau) \tag{6.3}$$

$$\hat{S} = W^{-1}(Z) \tag{6.4}$$

The schematic illustration of the steps performed in speech denoising using ASST is given in figure 6.3.

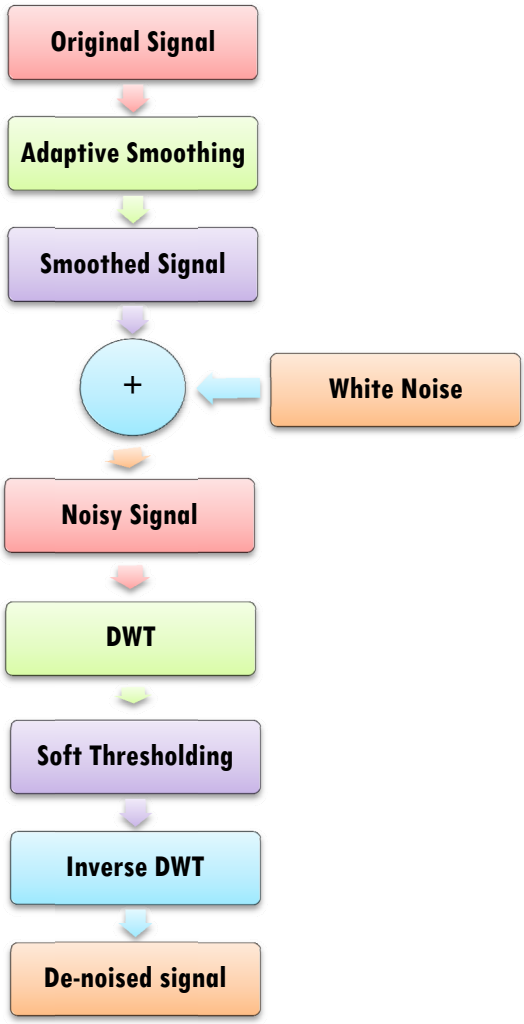


Figure 6.3: Schematic illustration of wavelet denoising using ASST

The implementation procedure is as follows. The speech signals captured are first smoothed using the proposed adaptive smoothing technique and then wavelet denoising based on soft thresholding is applied to the signal. There is an optimal decomposition level for denoising speech signals using

wavelet thresholding algorithms [162] These signals are then applied to DWPD algorithm for extracting features since they produced better results than the other methods. The dimensions of the feature vector set obtained are then reduced using PCA and are classified using ANN. These methods are implemented on the Malayalam words database created.

The main objective of selecting the wavelet, level and threshold is to maximise the SNR value and to minimise reconstructed error variance. The reconstruction quality increases for wavelets with more vanishing points.

6.6 Simulation Experiments and Results of ASST

For evaluating the performance of this method, the first step is to choose the wavelet family and the level of decomposition. For this, performance of different Daubechies wavelets of orders db8, db12, db20 and db22 are tested. Since better results were obtained using db20 at decomposition level 4, these were chosen for wavelet denoising. The speech signals are corrupted by white noise with SNR = 5dB and are used for evaluating the performance of the proposed algorithm. The performance analysis of the proposed adaptive smoothing algorithm is evaluated using SNR values, spectrograms and waveform plots.

6.6.1 Evaluation using SNR

The Signal-to-Noise Ratio (SNR) is a significant feature in determining the quality of audio data since recognition performance is strongly influenced by the SNR. It is a performance measure which is used to compare the level of a desired signal to the level of background noise which is represented as the ratio of signal power to the noise power. Actually it gives a measure of the actual useful information and unwanted or irrelevant data. SNR can be an important factor for determining reliability of a sub-band under noise

conditions [163, 164]. Achieving good SNR is a difficult task in a real world application since the signals are always affected by different types of noise. SNR can be calculated as

$$SNR_{out} = 10\log_{10}\frac{\sum s^2(n)}{\sum\left(s(n)-s'(n)\right)^2} \tag{6.5}$$

where *s(n)* is the original speech signal and *s'(n)* is the reconstructed signal.

SNR is used to quantify the amount by which the signal has been corrupted by noise and is calculated using the equation 6.5. Denoising is considered to be successful if the SNR value after denoising is greater than the SNR value before denoising. Table 6.2 given below shows the comparison of SNR values using Soft Thresholding (ST) alone and using Adaptive Smoothing Soft Thresholding (ASST). The SNR value obtained after denoising using db22 is less than that of denoising using db20. From the results obtained, it is clear that better results are obtained using db20.

Table 6.2: Comparison of SNR values using ST and ASST

Input SNR (db)	Wavelet used	Using ST	Using ASST
5	db8	6.974	7.797
5	db12	7.504	8.334
5	db20	8.263	9.437
5	db22	7.916	8.920

6.6.2 Evaluation using Spectrograms

Spectrogram gives the frequency domain representation of a speech signal. It is a better representation domain than waveform format which provides a visual representation of an acoustic signal. It shows the change in amplitude spectra over time and is represented using three dimensions where X-axis represents time (ms), Y-axis represents frequency and Z-axis colour

intensity which represents magnitude. In this representation, the complete speech sample is split into different time-frames and for each time- frame, the short-term frequency spectrum is calculated. The spectrogram provides a good visual representation of speech and can recognise distinctive patterns [165] and it also exhibits the dynamic changes in a speech spectrum. Figure 6.4 given below shows the spectrogram of the original signal, noisy signal, reconstructed signal using ST and reconstructed signal using ASST using db20.

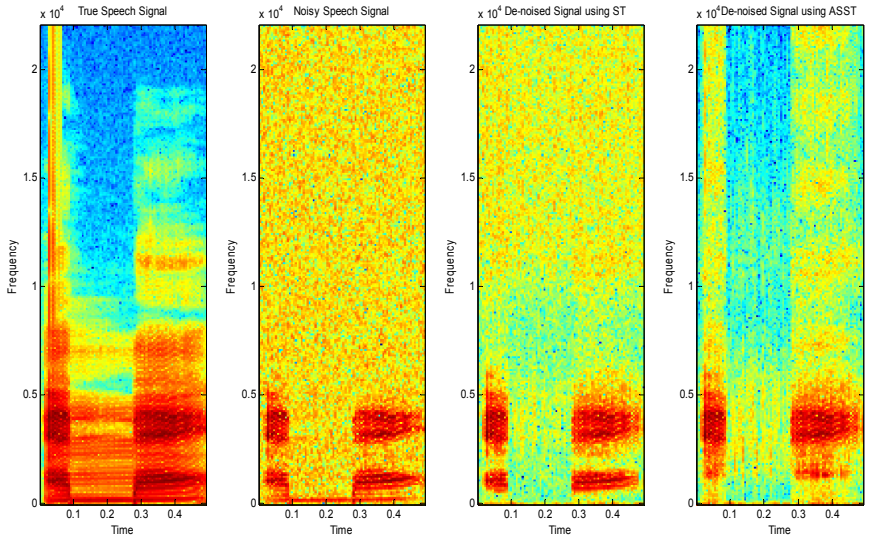


Figure 6.4: Spectrogram of original signal, noised signal with 5dB noise, denoised signal using ST and denoised signal using ASST

From the spectrograms it is clear that the original signal and the denoised signal using ASST are more similar. This obviously reduces error rate.

6.6.3 Evaluation using Waveform Plots

Waveform is the general form of representing a signal [9, 15, 88]. It represents the change of intensity with time and can be taken as the time domain representation of a speech signal. Sound intensity is often quoted in

deciBels (dB). The main disadvantage of waveform representation is that it also contains irrelevant data along with useful information. Figure 6.5 shows the waveform plot of the original signal, noisy signal, reconstructed signal using ST and reconstructed signal using ASST using db20.

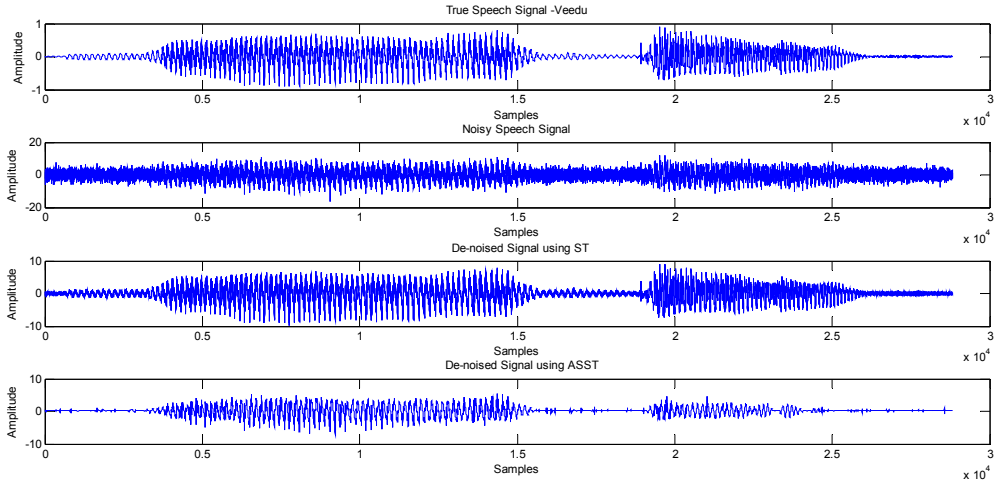


Figure 6.5: Waveform plot of original signal of word 'veedu' വീട്, noised signal with 5dB noise, denoised signal using ST and denoised signal using ASST

From the above evaluation techniques employed, it is clear that there is improvement in the SNR value and reduction in noise in the spectrogram and waveform plots. The results show the advantages of using adaptive smoothing for the enhancement of the signals.

6.7. Experimental Results of the Speech Recognition System

The smoothed signals after feature extraction using DWPD gnerates a feature vector set which is then classified using ANN. A recognition accuracy of 92.66% is obtained after classification which is better than that of the results obtained without performing adaptive smoothing. The figure 6.6 shows the confusion matrix of the results obtained after classification using ASST.

Predicted Class

Actual Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	975	0	0	0	4	0	3	0	8	1	0	5	0	2	0	1	1	0	0	0
2	7	906	1	7	2	19	0	6	3	6	0	6	12	0	6	0	7	0	8	4
3	0	1	997	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
4	10	12	23	796	0	20	4	14	18	12	1	6	6	42	0	6	13	8	7	2
5	0	0	16	1	895	0	1	14	0	0	1	5	16	15	0	5	1	11	0	19
6	1	17	0	8	0	898	0	11	5	2	0	15	0	0	13	0	20	0	10	0
7	0	0	1	0	0	0	995	0	0	2	0	0	1	0	0	0	0	1	0	0
8	0	0	0	2	0	0	0	995	0	0	0	1	0	0	0	0	2	0	0	0
9	1	0	4	0	0	8	0	5	957	0	0	7	0	1	7	0	0	4	0	6
10	8	0	2	10	11	0	1	0	23	905	0	0	15	1	12	0	4	0	8	0
11	14	0	24	2	2	0	0	11	0	0	915	0	0	9	0	1	15	0	1	6
12	16	0	10	15	0	13	0	8	19	0	3	866	0	13	3	0	7	4	22	1
13	0	0	3	0	0	0	0	1	0	0	0	0	995	0	0	1	0	0	0	0
14	0	5	11	23	30	0	20	11	26	11	23	0	10	750	0	12	16	8	22	22
15	0	11	0	5	0	1	0	0	4	0	3	0	4	0	964	0	4	0	2	2
16	0	0	0	0	0	0	0	1	0	0	0	0	0	0	4	995	0	0	0	0
17	2	15	11	0	0	12	0	0	7	0	8	7	0	5	0	1	927	0	5	0
18	0	0	0	2	0	0	0	0	0	0	2	1	0	1	0	0	0	994	0	0
19	0	0	1	0	0	0	0	0	0	0	0	0	1	0	2	1	0	0	995	0
20	10	13	1	11	10	8	20	0	11	13	1	10	0	32	7	12	3	8	19	811

Figure 6.6: Confusion matrix obtained using smoothing and DWPD

6.8 Comparison of Results using ST and ASST

In this section the results obtained using ST alone and ASST are evaluated. Table 6.3 shows the classification results obtained using ST and ASST.

Table 6.3: Comparison of classification results using ST and ASST

Pre-processing Method	Precision	Recall	Correctly Classified	Accuracy %
ST	0.9	0.895	17901	89.505
ASST	0.93	0.927	18531	92.66

The results obtained clearly shows that smoothing improves the recognition rate.The graph given in figure 6.7 shows the comparison of the results obtained using DWPD method and DWPD method after smoothing.

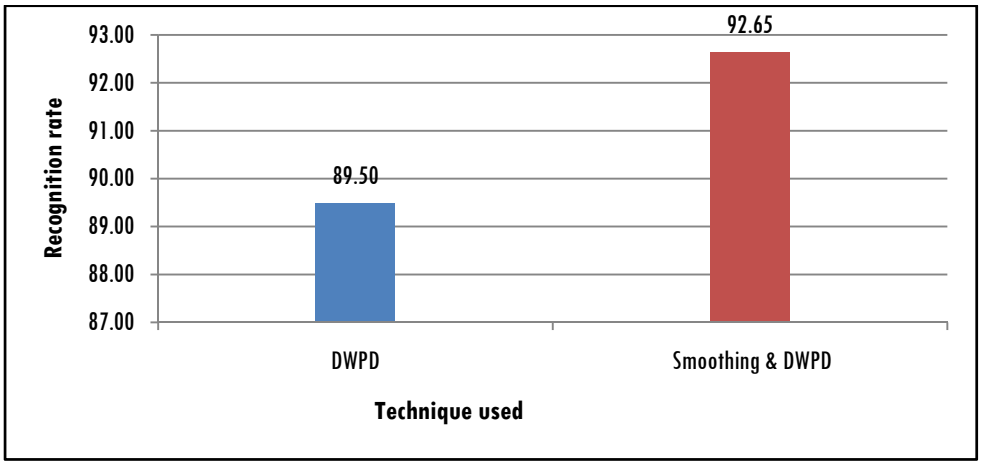


Figure 6.7: Comparison graph of DWPD and DWPD after smoothing

6.9 Summary of the Chapter

The performance of a speech recognition system improves when noise present in the signals are removed. So in this chapter, a new algorithm is developed for removing the sudden spikes in the signal thereby smoothing the signal. The new algorithm is implemented on the speech signals during the pre-processing stage. Then a hybrid method is devised by passing the smoothed signal after adaptive smoothing to wavelet denoising using Soft Thresholding. Then these smoothed and denoised signals are applied to the DWPD method for feature extraction and are finally classified using MLP. The results obtained clearly show the need for removing noise from the speech signals for better recognition of speech. The new algorithm for adaptive smoothing of the signal is developed during the pre-processing stage. The two forthcoming chapters will therefore deal with the modifications and new proposed methods for post processing and classification.

Chapter 7

STATISTICAL THRESHOLDING TECHNIQUES FOR POST PROCESSING

Contents

In the preceding two chapters, two new algorithms were developed- namely DWPD which was designed for obtaining more relevant features and ASST for removing noise from the signals by smoothing. The third module in developing a speech recognition system is the post processing module. This chapter suggests three statistical thresholding techniques to be applied during the post processing stage based on Three Sigma Limits, Quartiles and Confidence Interval Mode. These are applied to the feature vector set obtained from the DWPD method in order to bring the feature values to a more consistent and polished format. The feature vectors when applied with these statistical thresholding techniques generates better results during classification.

7.1 Introduction

The selection of the most relevant features from the feature vector set generated is essential for the proper and efficient classification of the speech samples. The feature vectors generated may have disparity in the feature vectors obtained due to alteration in the signals. This may be due to the

adverse environmental conditions or variability of speech related to differences in the vocal tract among different speakers. So significant efforts should be carried out for transforming the feature vectors obtained to reduce these effects during post processing stage [153]. During post processing, the influence of the irrelevant and useless features are removed by applying different thresholding techniques. This significantly improves the comprehensibility which in turn improves the performance of the system during classification.

From an engineering point of view, feature selection process takes place during post processing. Feature selection is performed to select the relevant and informative features for further classification. This also includes transformation to a relatively low dimensional feature space by preserving the information pertinent to the application called feature set reduction, performance improvement techniques for improving the accuracy, transformation of the feature vector set to a format which is more compatible for classification etc [166]. If the number of features is more, the dimension of the feature vector set increases and this imposes severe need for computation as well as storage during training and testing [167].

In the previous chapters, two techniques were employed during the post processing stage namely Normalisation and Principal Compound Analysis. Normalisation was used to bring the data values between a range and PCA was applied in order to reduce the dimensions of the feature vector set. In this chapter, 3 techniques based on statistical thresholding methods are applied namely Three Sigma Limits, Quartiles and Confidence Interval Mode to the feature vector set obtained. The performance of these three techniques in selecting the relevant speech features are tested and evaluated.

The chapter is organised as follows. Section 7.2 presents a brief introduction to the statistical thresholding methods and a brief description of the three statistical thresholding techniques used in this work. The implementation algorithms for these three thresholding techniques are explained in the next section. Section 7.4 presents the experiments performed using these three techniques and the results obtained. A comparison of results is performed in the subsequent section. The chapter concludes with section 7.6.

7.2 Statistical Thresholding

Thresholding techniques are used to limit the set of values of the features below a threshold value or to limit the values of the features within a certain range. This range is defined differently depending on the central value we take. But the actual data values may include values outside this predefined range.

In this work, three thresholding techniques based on statistical distribution methods, namely Three Sigma Limits, Quartiles and Mode based thresholding have been used. Instead of selecting one value for the threshold limit, two limits are used - Upper Specification Limit (USL) and Lower Specification Limit (LSL) [168]. These are used to limit the values of the feature set to a uniform format so that the recognition rate can be improved. These techniques are based on replacing the data values outside the upper and lower limits with mean, median and mode respectively. The Mean, Median and Mode are all valid measures of central tendency which represents a single value that attempts to describe a set of data by identifying the central position within that set of data. But depending on the different conditions, some measures of central tendency become more appropriate to use than others. The three statistical thresholding techniques that are implemented in this research work are explained in the following section.

7.2.1 Three Sigma Limits

This is a statistical calculation that is used to refer to data within three standard deviations from a mean. Usually 3 sigma limits are used to set the upper and lower control limits in statistical quality control charts. Here σ represents standard deviation in statistical analysis and μ denotes the mean which are the fundamental building blocks in Statistics. Standard deviation is a measure of how flat a data distribution is. High value of sigma indicates that the data is more dispersed from the norm.

When lower and upper limits are established, it indicates that the data points under and over it need a significant change. In Statistical Process Control (SPC), it is stated that the control limits placed at three standard deviations from the mean in either direction provide an economical tradeoff between the risk of reacting to a false signal and the risk of not reacting to a true signal - regardless the shape of the underlying process distribution [169]. In a normal distribution, about 68.27% of all the values in the set are within one standard deviation of the set's norm, about 95.45% of the values lie within 2 standard deviations of the mean and nearly 99.73% of the values lie within 3 standard deviations of the mean [170]. Three sigma limits is a statistical rule which dictates that for a normal distribution, almost all values fall within 3 standard deviations of the mean value. Here, the Three Sigma Limits of each feature is calculated and the values outside the limits are replaced by the mean value of that feature which in turn ensures process stability.

7.2.2 Quartiles

It is a method which is used in descriptive Statistics. For calculating the Quartiles, the set of values are divided into four groups using three points. Each quarter has the same number of observations [171]. The first Quartile Q_1

splits the data in the ratio 1:3. The second Quartile Q_2 is the median which divides the data set into two equal parts (1:1). The upper Quartile Q_3 splits the data in the ratio 3:1. The difference between the upper and lower quartiles is called the inter quartile range. Median is calculated by cutting the data into two groups with equal number of points and then by taking the middle value that separates these groups [172]. In this work, first the three quartiles are defined for each feature. If the value of the feature is greater than Q_3 or less than Q_1, then this value is replaced by the median of the corresponding feature values. Quartiles can be calculated in different ways depending on the number of data values. The commonly used methods are [172, 173, 174]:

1. Divide the ordered data set into two halves using the median without including the median into the halves. Then the lower quartile value becomes the median of the lower half of the data and the upper quartile value becomes the median of the upper half of the data.
2. Use the median to divide the ordered data set into two halves. If the median is a datum rather than the mean of the middle two data, include the median in both halves. Now, lower quartile value becomes the median of the lower half of the data. The upper quartile value is the median of the upper half of the data.
3. If there is an even number of data points, then the method is the same as above. If there are $(4n+1)$ data points then,
   - The lower quartile is calculated as 25% of the n^{th} data value plus 75% of the $(n+1)^{th}$ data value and the upper quartile is 75% of the $(3n+1)^{th}$ data point plus 25% of the $(3n+2)^{th}$ data point.
   - If there are $(4n+3)$ data points, then the lower quartile is 75% of the $(n+1)^{th}$ data value plus 25% of the $(n+2)^{th}$ data value and the upper

quartile is 25% of the $(3n+2)^{th}$ data point plus 75% of the $(3n+3)^{th}$ data point.

7.2.3 Confidence Interval Mode

Mode is the most frequently occurring value in a set of measurements which is the value at which the peak of the distribution curve occurs. Mode is a measurement of relatively great concentration. For a given data set, there can be more than one mode. In statistical distribution, if there is only one value for the mode, then it is called unimodal distribution. If there are two modes then it is called bimodal and more than 2 modes are called multimodal. A distribution in which each different measurement occurs with equal frequency is said to have no mode. The sample mode can be considered as the best estimate of the population mode. A mode is affected more by sampling and grouping than skewness, but mean and median are more affected by skewness [175]. In a unimodal distribution, the difference between the mean and the mode is represented as

$$\frac{|\ mean - mode\ |}{standard\ \ deviation} \leq \sqrt{3} \tag{7.1}$$

The same bound can be used to find the difference between the mode and the median. This can be expressed as

$$\frac{|\ median - mode\ |}{standard\ \ deviation} \leq \sqrt{3} \tag{7.2}$$

Here, the range is chosen as $\mu - \sqrt{3}\sigma$ and $\mu + \sqrt{3}\sigma$ and the values outside this range are substituted by the mode of that particular feature.

7.3 Implementation

The smoothed signals after performing ASST are applied to the DWPD method for extracting the features. Statistical thresholding techniques

described in the above section are applied to this feature vector set. This brings the feature vectors into a more compact and consistent format. This section explains the implementation algorithms involved in the feature selection procedure during post processing using the above given three statistical thresholding methods.

7.3.1 Implementation using Three Sigma Limits

In this method we have seen that the feature values that are outside the Three Sigma Limits are substituted by the Mean. Arithmetic mean is the most widely used measure of central tendency. In this work, the mean is calculated as the sum of all data values of a feature divided by the number of observations in that particular feature. Table 7.1 describes the algorithm for post processing using Three Sigma Limits.

Table 7.1: Steps for post processing using Three Sigma Limits

1. For each feature do the following

   1.1 Find the mean μ of that feature obtained after feature extraction using the equation

   $$\mu = \frac{\Sigma Xi}{n}$$

   1.2 Calculate the standard deviation σ of the feature using the equation

   $$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

2. For each observation X in a feature, do the following.

   If $\mu - 3\sigma < X < \mu + 3\sigma$ *then* $X = X$

   Else if

   $X > \mu + 3\sigma$ *or* $X < \mu - 3\sigma$ *then* $X = \mu$

7.3.2 Implementation using Quartiles

In this method, the data values of the feature which are outside the first Quartile Q_1 and upper Quartile Q_3 are replaced by Median. Median gives the middle measurement in an ordered set of data. In this work, first the data values in each feature are arranged in ascending order. If n is odd, then the Median is calculated as

$$Median = X_{(n+1)/2} \tag{7.3}$$

where n is the total number of observations in a particular feature. If n is even, then the subscript will be a half integer which indicates a number between two integers. Then the median is calculated as the midpoint between them. The algorithm for post processing the feature vectors obtained using quartiles is given in Table 7.2.

Table 7.2: Steps for post processing using Quartiles

1. For each feature do the following
   - *1.1 Order the values of each feature from smallest to largest.*
   - *1.2 Find the first quartile Q_1, second quartile Q_2 and third quartile Q_3 of each feature*
   - *1.3 Calculate the median of each feature using the equation 7.3*
2. For each observation X in a feature, do the following

   If $Q_1 < X < Q_3$ *then* $X = X$

   Else if

   $X > Q_3$ *or* $X < Q_1$ *then* $X = Median(X)$

7.3.3 Implementation using Confidence Interval Mode

Here, the data values of a feature which are outside the defined range are replaced by the mode which returns the most frequently occurring data value among the set of feature values. The implemtation algorithm for confidence interval mode is given in table 7.3.

Table 7.3: Steps for post processing using Mode

1. For each feature do the following

1.1 Find the mean μ of that feature obtained after feature extraction using the equation

$$\mu = \frac{\Sigma Xi}{n}$$

1.2 Calculate the standard deviation σ of the feature using the equation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

2. For each observation X in a feature, do the following.

If $\mu - \sqrt{3}\sigma < X < \mu + \sqrt{3}\sigma$ then $X = X$

Else if

$X > \mu + \sqrt{3}\sigma$ *or* $X < \mu - \sqrt{3}\sigma$ then $X = mode(X)$

7.4 Experiments and Results

After performing feature selection using the above statistical thresholding techniques, the feature sets are given to MLP for classification. The results obtained using these methods are given below.

7.4.1 Result Analysis using Three Sigma Limits

In this method, the values which are outside the range of $\mu - 3\sigma$ and $\mu + 3\sigma$ are replaced by the mean of the values of a particular feature. MLP produced an accuracy of 93.77% which is better than the earlier method. The confusion matrix given in figure 7.1 explains the results obtained using Three Sigma Limits.

Predicted Class

Actual Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	959	0	0	3	0	5	0	4	0	16	0	8	1	0	0	3	1	0	0	0
2	6	724	0	5	9	14	0	35	8	17	0	8	97	0	28	25	1	1	13	9
3	0	1	974	2	0	1	1	0	0	0	0	0	2	0	0	5	12	0	0	2
4	7	2	15	903	14	2	0	0	5	14	0	1	0	13	2	3	0	10	0	9
5	0	4	3	2	954	0	6	1	0	2	1	2	0	1	1	3	12	0	8	0
6	2	5	0	7	0	951	0	1	6	3	0	2	7	0	2	4	1	5	0	4
7	0	3	1	1	0	1	985	0	0	0	2	2	1	0	1	3	0	0	0	0
8	3	7	1	0	0	0	0	985	0	0	0	0	1	0	2	1	0	0	0	0
9	3	0	4	0	0	2	0	1	957	2	0	18	1	0	1	7	1	0	3	0
10	8	1	0	11	9	2	0	0	2	940	0	1	0	0	3	0	4	0	4	15
11	8	1	0	1	10	1	3	12	1	0	896	0	2	20	3	4	7	15	1	15
12	15	0	2	0	1	1	8	17	5	17	0	901	0	11	0	10	0	4	0	8
13	2	0	4	0	6	1	0	13	0	0	0	3	955	0	4	4	1	0	7	0
14	1	0	1	3	13	1	0	2	0	1	3	0	8	962	0	3	1	1	0	0
15	0	6	0	0	8	0	0	4	0	2	0	6	2	5	957	4	0	0	5	1
16	1	5	9	19	10	1	18	0	9	0	9	0	11	1	20	866	2	0	10	9
17	0	6	1	1	0	1	0	12	0	9	15	0	0	0	3	1	940	0	0	11
18	2	3	1	1	0	2	0	0	0	0	0	0	0	0	0	4	0	986	0	1
19	0	2	0	0	2	1	0	1	1	0	0	0	0	0	3	3	0	0	986	1
20	0	0	0	5	0	1	1	0	0	5	0	2	1	1	2	9	0	0	0	973

Figure: 7.1: Confusion matrix obtained using Three Sigma Limits

7.4.2 Result Analysis using Quartiles

Here, the values outside the quartiles Q_1 and Q_3 are replaced by the median of that feature. The feature vector set obtained when classified using MLP, produces a recognition accuracy of 95.78%. So the results obtained using Quartiles are found to be better than the results obtained using Three Sigma Limits. Figure 7.2 presents the detailed classification results obtained using the confusion matrix.

Actual Class \ Predicted Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	995	1	0	2	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
2	0	973	0	4	0	5	0	0	1	0	1	2	1	0	4	0	6	0	2	1
3	0	0	995	0	1	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0
4	1	0	1	986	0	0	4	0	1	5	1	0	0	0	0	0	0	0	0	1
5	0	0	0	0	995	0	0	0	0	0	0	2	0	0	0	1	0	0	0	2
6	0	0	0	0	1	995	0	0	0	0	1	0	0	0	0	1	0	2	0	0
7	0	1	0	0	0	0	989	0	0	0	3	0	0	0	0	0	6	1	0	0
8	0	0	0	1	1	0	0	995	1	0	2	0	0	0	0	0	0	0	0	0
9	0	4	0	1	1	0	0	2	985	0	0	1	5	0	0	0	1	0	0	0
10	1	0	0	0	0	1	0	0	0	995	0	0	0	0	0	1	0	2	0	0
11	0	2	0	2	0	0	4	0	0	0	983	3	0	0	0	0	1	1	0	4
12	0	0	0	0	0	2	0	0	1	0	0	995	0	0	0	0	1	0	0	1
13	0	2	0	0	0	0	0	0	9	0	1	0	984	0	0	0	0	3	0	1
14	6	15	0	10	0	1	6	0	24	0	2	0	19	895	0	14	1	0	6	1
15	11	0	5	1	0	3	0	7	3	0	1	8	6	0	941	0	6	1	7	0
16	10	0	5	0	4	1	0	3	7	0	11	0	1	0	3	948	1	1	0	5
17	14	8	15	6	20	41	22	14	43	18	1	22	6	35	5	3	697	18	7	5
18	0	0	0	1	1	0	0	0	6	0	5	0	0	0	0	0	0	986	0	1
19	0	12	0	8	1	1	0	8	1	0	2	0	11	0	4	0	16	0	928	8
20	0	2	1	10	0	0	12	0	7	0	18	0	13	11	0	0	29	1	0	896

Figure: 7.2: Confusion matrix obtained using Quartiles

7.4.3 Result Analysis using Confidence Interval Mode

In this method, the feature values that are outside μ - $\sqrt{3}\sigma$ or $\mu + \sqrt{3}\sigma$ are replaced by the mode of the feature. Then the new feature vector set obtained are applied to the MLP structure of the ANN which produced an accuracy of 92.815. Thus it is evident from the results obtained that the recognition rate generated using confidence interval mode is less than that of the recognition accuracy generated by Three Sigma Limits and Quartiles. The confusion matrix given in figure 7.3 depicts the detailed results obtained using confidence interval mode.

Predicted Class

Actual Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	965	0	4	6	0	1	0	0	0	7	2	0	1	0	7	1	1	5	0	0
2	3	897	2	2	4	1	11	14	1	8	8	7	1	7	1	12	1	10	8	2
3	0	3	946	6	2	0	1	5	8	0	5	0	1	10	0	2	0	1	9	1
4	3	2	0	979	0	0	0	1	1	1	1	0	0	1	1	2	0	0	1	7
5	0	2	10	0	910	1	10	1	6	12	3	4	1	5	15	1	8	1	0	10
6	2	3	1	0	0	986	0	1	1	0	0	0	2	0	2	0	1	0	1	0
7	1	2	1	1	0	0	985	0	1	1	1	0	0	0	1	3	0	0	0	3
8	2	5	2	0	0	1	0	985	0	0	0	0	1	0	2	1	0	1	0	0
9	8	0	1	0	0	6	0	4	951	3	0	2	0	5	0	4	1	6	0	9
10	10	1	6	5	0	0	0	0	2	960	2	0	0	1	0	0	2	0	0	11
11	6	1	2	1	13	2	10	1	5	10	911	0	2	12	0	2	14	5	2	1
12	9	0	7	0	3	1	10	10	13	0	20	878	5	0	18	11	5	1	1	8
13	2	23	10	3	30	1	14	2	25	21	2	5	772	10	5	16	20	28	1	10
14	2	3	1	2	1	8	8	6	4	6	11	6	10	912	3	1	3	2	7	4
15	0	6	1	3	4	1	13	15	5	5	3	7	2	26	897	1	1	3	2	5
16	1	2	8	2	3	2	0	2	4	0	2	6	1	1	3	960	0	3	0	0
17	10	11	2	8	6	5	3	19	16	1	7	1	8	15	1	2	866	15	3	1
18	2	2	2	1	12	3	3	2	16	8	7	14	2	5	4	4	7	887	4	15
19	0	7	2	0	0	0	0	2	0	0	1	1	0	0	1	0	0	1	985	0
20	8	5	2	2	4	9	5	2	0	1	2	4	2	1	6	1	9	2	5	930

Figure: 7.3: Confusion matrix obtained using Confidence Interval Mode

7.5 Comparison of Results

From the three experiments performed during post processing stage, it is clear that there is notable improvement in all the three methods. In the previous chapter deploying adaptive smoothing and DWPD without using statistical thresholding methods, recognition rate obtained was 92.66%. Among the three thresholding methods, the results obtained using quartiles outperformed others. Table 7.4 shows the comparison of results obtained using ASST (without statistical thresholding methods) and results obtained using the statistical thresholding techniques, Three Sigma Limits, Quartiles and Confidence interval mode.

Table 7.4: Performance evaluation of statistical thresholding techniques

Post processing Method	Precision	Recall	Correctly Classified	Accuracy %
Without thresholding	0.93	0.927	18531	92.66
Three sigma limits	0.938	0.938	18754	93.77
Quartile	0.964	0.958	19156	95.78
Mode	0.933	0.928	18563	92.815

The figure 7.4 given below shows the comparison of results obtained using ASST and DWPD without thresholding, three sigma limits, quartiles and mode.

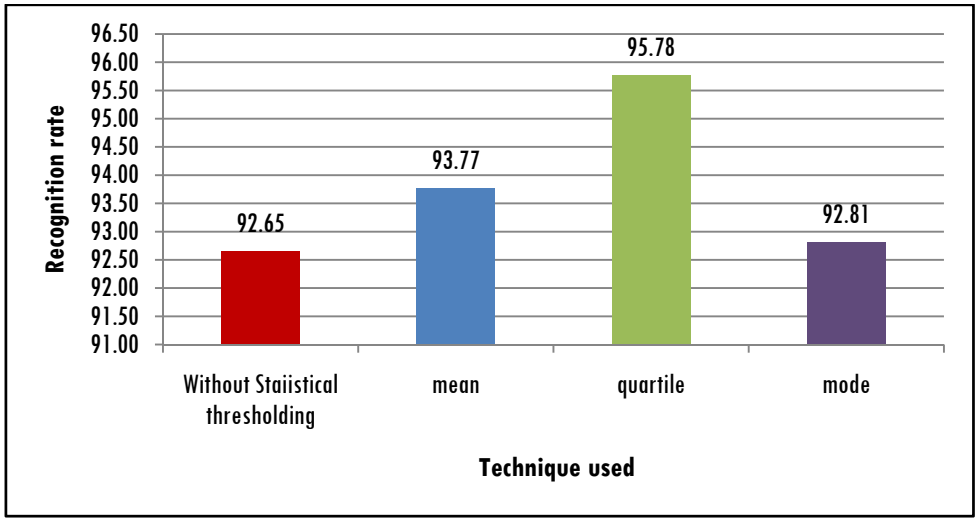


Figure 7.4: Graph showing the comparison of statistical thresholding techniques

When compared to mean, the median expresses less information since it does not take into account the actual value of each measurement. Median takes into consideration only the rank of each measurement. But, for some applications it is found to be better. One of the advantages of median is that it is not affected by the extreme high and extreme low measurements. But mean is affected by these values. So, for skewed data, median is more preferable to mean to express the central tendency. Though Mode is affected by skewness less than that of the mean or the median, it is more highly affected by sampling and grouping than the other two measures [175]. In this study, the performance of quartiles based on median is found to be superior to Three Sigma Limits based on mean and confidence interval mode. This is due to the fact that the skewness generated for each feature is found to be high when the data analysis is done using Descriptive Statistics in Microsoft Excel.

7.6 Summary of the Chapter

This chapter discusses the role of post processing techniques for the proper selection of the feature vectors. This exploits the capabilities of statistical thresholding techniques which utilises the properties of statistical distributions like mean, median and mode for the efficient recognition of speech signals. This also demonstrates the improvements obtained in the recognition rate by using these three statistical thresholding techniques. The three techniques used are found to perform well for processing the feature vectors obtained. The feature vectors selected using quartiles generated better results. By incorporating the statistical techniques into signal processing, better results are obtained. So, it is concluded that post processing steps also play an important role in improving the performance of the speech recognition system since good features are essential for better classification.

After post processing, the last and final stage in the development of a speech recognition system is the classification stage, which is the back end processing stage of the speech recognition system. So studies are conducted to improve the recognition rate obtained during the classification stage in the next chapter.

Chapter 8

ENSEMBLE CLASSIFICATION AND PERFORMANCE EVALUATION OF RESULTS

Contents

In the previous chapters, we had already developed different speech recognition systems by designing new algorithms and methods during pre-processing, feature extraction and post processing stages. The fourth and the final stage of development of the speech recognition system is the classification stage. Here, the classifier models are combined together using ensemble classification techniques to produce better performance than a single base level classifier model. This chapter also presents a comparison of the performance evaluation of all the speech recognition systems developed in this research work.

8.1 Introduction

Pattern recognition is the scientific discipline of classifying patterns into a set of categories called classes [165]. Classification tasks require the construction of statistical models that represent mapping from input data to the appropriate outputs. While experimenting with a classifier, there are many parameters that affect the performance of the classifier. When different training parameters are tried for the same classifier, individual classifiers are

allowed to generate different decision boundaries. This produces different errors that are generated for the same classifier. By choosing a proper combination of these parameters, the total error produced by each classifier can be reduced [94]. This is the case with a single classifier. However, when different classifiers are combined with suitable training parameters, the performance can be improved further.

Research in Data Mining and Pattern Recognition approaches have led to the combining of different classifiers instead of experimenting with a single classifier. The different pattern classifiers used with ASR systems may have benefits and shortcomings [176]. So, recognition of speech by using only a single classifier is not considered to be efficient. In order to overcome the limitations of a single classifier, different classifiers can be combined in specific formats for the better performance of the speech recognition system. The main objective of this part of the work is to test whether a combined classifier model performs better than a single classifier by improving the overall performance of the speech recognition system developed and if so which combination is most suitable for this work. For this, different ensemble classification techniques are employed and a comparison of these techniques in classifying the speech features is made.

In chapters 3 and 4, we had experimented with four classifiers namely MLP, HMM, Naive Bayes and SVM for recognising the speech samples in Malayalam along with the feature extraction techniques such as LPC, MFCC, DWT and WPD. Among these classifiers, the results obtained using MLP classifier was found to be better than that of the other classifiers and further experiments were done using MLP classifier. In this chapter, the performance of three popular ensemble classifier techniques are analysed and evaluated namely Bagging, Boosting and Stacking. The different classifiers which are

used for ensemble learning are MLP structure of ANN, SVM and Naive Bayes classifiers since they produced better results. Moreover, HMM was not found to be efficient during ensemble learning.

The chapter is organised as follows. Section 8.2 provides the basic concepts of ensemble learning. In the next section, a brief description of the popular ensemble learning algorithms that are applied in this research work are explained. Section 8.4 presents the implementation procedure for combining the classifiers using the different ensemble techniques. A comparison of the results obtained is given subsequently. Finally, a comparison of the performance evaluation of all the speech recognition systems developed in this research work using different pre-processing, feature extraction, post processing and classification methods are discussed in section 8.6. The chapter is concluded in section 8.7.

8.2 Ensemble Learning Concepts

Ensemble learning is a rather new concept which is used to improve the classification results by combining outputs of different classifiers. It is the process of combining multiple models, for example classifiers, to solve a particular pattern recognition problem for improving the performance of the model. It is also used to assign a confidence to the decision made by the model by selecting optimal features, data fusion, incremental learning and error-correction [177]. These are also known as multiple classifier systems because it is a combination of different classifiers. An ensemble is a supervised learning algorithm which can be trained and tested for classifications and predictions. In ensemble learning, several classifiers are aggregated whose individual predictions are combined in some manner, may be by voting or averaging, to form the final prediction [178]. The main motivations of using ensemble classification are to:

a) *reduce variances and*

b) *reduce bias.*

This is due to the fact that multiple classifiers can learn more than a single classifier and the results obtained are less dependent on a single training set. Bias, variance and total error can be calculated as

Bias = expected error of the combined classifier on new data

Variance = expected error due to the particular training set used

Total expected error = bias + variance

By reducing the variance, results are less dependent on peculiarities of a single training set and by reducing the bias, a combination of multiple classifiers may learn a more expressive concept class than a single classifier. From statistical, computational and representational point of view, ensemble classifiers are considered to be better than a single classifier [179]. The generalisation ability of an ensemble is better than that of individual classifier or learner present in an ensemble classifier. Generally, for getting an efficient ensemble, the base learners should be accurate and diverse as possible [149, 180].

Each ensemble method has different properties that make it better suited to a particular type of classifier and/or application. Ensemble methods are designed to combine multiple classifiers to improve robustness as well as classification performance from any of the constituent classifiers. It can make use of a 'divide and conquer approach' where a complex problem is decomposed into multiple sub-problems that are easier to understand and solve [181]. The main criterion for the success of the ensemble approach is the diversity in the individual classifiers with respect to misclassified instances [177] and this diversity can be achieved by using:

a) *different training data to train single classifiers,*

b) *different training parameters,*

c) *different features to train the classifiers and*

d) *combination of different types of classifiers*.

8.3 Ensemble Learning Methods Applied in this Research Work

There are different ways to combine the classifier models for a better classification approach based on the scheme chosen and the type of the classifier and there exists different ensemble classifying methods. There are various factors that differentiate between these methods like inter-classifier relationship, combining method, ensemble size and diversity generator [182]. The most commonly used approaches are based on two schemes namely *Voting and Stacking.* Among the various techniques available, *Bagging and Boosting* [183] are the two popular methods which are based on the voting scheme. This is similar to the voting conducted in our daily life for selecting a candidate, if there is a conflict in the opinion among the members of a group. Researches show that bagging ensemble generally produces better results whereas boosting generates widely varying results depending on the characteristics of the database [184]. In stacking, the predictions generated by each different classifier model are given as the input to the ensemble classifier and the output produced by this is taken as the final class. There are two types of ensemble classification based on the type of the classifier used. They are homogeneous ensemble where the classifiers used are of the same type and heterogeneous ensemble where the classifiers are of a different nature. Voting is the simplest way to combine the output of multiple classifiers within a voting framework and is found to be efficient [185].

There is no single learning algorithm that in any domain always induces the most accurate learner. Each learning algorithm dictates a certain model with a set of assumptions, leading to the corresponding model bias. If the assumptions do not hold for the data, the model bias leads to error. We construct a group of base learners which, when combined, has higher accuracy than the individual learners. The base learners should be accurate on different instances, specialising in different sub domains of the problem, so that they can complement each other. The base learners work in a parallel pattern. Given an instance, they all give their decisions which are then combined to give the final decision. The main advantage of using combination techniques is to achieve results superior to the single best classifier. This is based on the assumption that the errors made by each of the classifiers are not identical and if we combine multiple classifier outputs in an efficient manner, we may be able to correct some of these errors [186]. The most popular ensemble learning methods namely Bagging, Boosting and Stacking are explained below.

8.3.1. Bagging

Bagging stands for **b**ootstrap **agg**regat**ing**. It is one of the earliest, simplest and most successful ensemble based algorithms which gives good performance [187]. In this method, different training data subsets are randomly chosen from the entire training dataset. Each training data subset is used to train a different classifier of the same type and is recorded [188, 189]. Individual classifiers are then combined by taking a simple majority vote of their decisions. The class chosen by most number of classifiers is selected [190]. This allows each base classifier to be trained with different random subset of the patterns. Bagging works well for unstable procedures like Neural Networks which may cause large variations in the output classifier even for a small change in the training data and is a relatively easy way to improve an

existing method. Figure 8.1 given below shows the schematic illustration of bagging ensemble learning method [191].

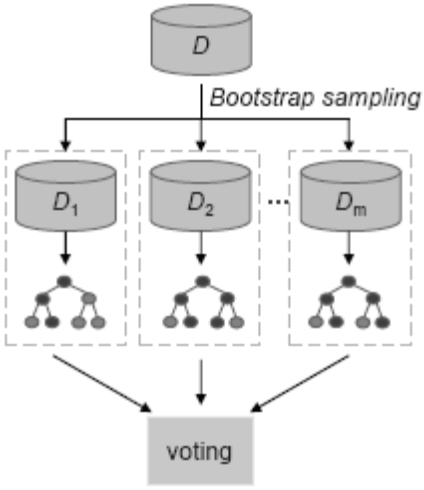


Figure 8.1: Schematic illustration of Bagging ensemble learning

Given a training set *D* of size *n,* generate a new training set *D1* of size *N* by sampling examples from *D* uniformly and with replacement, known as a bootstrap sample *m* training sets. Run the learning algorithm *m* times, each time with a different training set. The Bagged classifier then combines the predictions of the individual classifiers to generate the final outcome. This method is very useful for large and high-dimensional data, where it is difficult to find a good model or classifier due to the complexity and scale of the problem.

8.3.2. Boosting

Boosting is also similar to bagging since it also creates an ensemble of classifiers by re-sampling the data, which are then combined by majority voting [188,192]. But here, the construction of the model is different from that of bagging. In this method, the misclassified classifier models are allowed to

participate in the training process more number of times. Each classifier is associated with individual weights for their accuracies and the class having the maximum weight is assigned. Bagging is considered to be better than boosting because boosting suffers from the problem of over fitting - that is, it works well for training data but is not so good for unknown data. Another limitation of boosting is that it is applied only to binary classification problems. This limitation is overcome with the AdaBoost algorithm. In boosting, successive classifiers depend upon its predecessors by looking at errors from previous classifiers to decide what to focus on for the next iteration over data. The schematic illustration of boosting ensemble learning method [191] is given in figure 8.2.

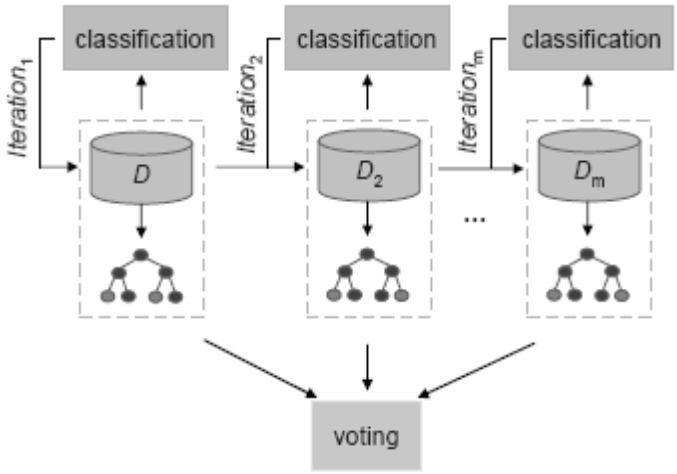


Figure 8.2: Schematic illustration of Boosting ensemble learning

8.3.3. Stacking

In stacking, the predictions obtained from different classifiers are given as input to a meta learner. This combines the predictions to create a final best predicted classification. Many researches have shown that by combining the predictions from multiple methods produce more accurate results than the

results that can be derived from any one method [193]. The main concept used in stacking is to use a new classifier to correct the errors of a previous classifier. A number of first-level individual learners are generated from the training data set by employing different learning algorithms. The individual learners are then combined by a second-level learner which is called as meta-learner. This will combine the predictions from the different models to yield maximum classification accuracy. In contrast to stacking, no learning takes place at the meta-level when combining classifiers by a voting scheme [181]. Figure 8.3 given below illustrates the schematic illustration of stacking ensemble method.

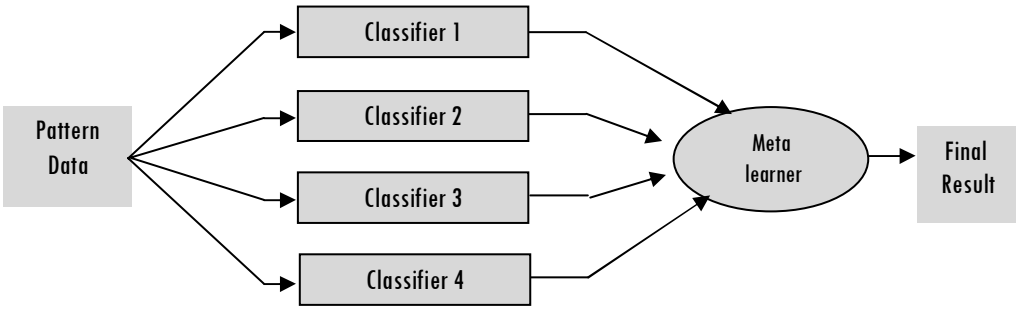


Figure 8.3: schematic illustration of of Stacking ensemble learning

8.4 Implementation

This section presents the implementation procedure for combining the classifiers using three popular ensemble techniques such as Bagging, Boosting and Stacking. In this work, three classifiers participate in the ensemble learning process namely ANN, SVM, and Naive Bayes classifier. The performance of these three classifiers were already evaluated in the previous chapters and among these classifiers, the best recognition rate was obtained using the MLP structure of the ANN classifier. The rest of this section explains the algorithms and the procedure for implementing these ensemble methods.

8.4.1 Implementation using Bagging Ensemble Framework

Researches in data mining and pattern recognition have shown that, voting classification algorithms like bagging and boosting are found to be very successful in improving the accuracy of certain classifiers. Bagging ensemble methods are one of the earliest and simplest ensemble based algorithms which are found to be good in improving the performance of unstable methods by scaling down the variance.

In bagging, the base learning algorithm is run repeatedly in a series of rounds. Here, in each round, the base learner is trained on a bootstrap replicate of the original training set. Suppose the training set consists of *m* examples. Then a bootstrap replicate is a new training set that also consists of *m* examples, and which is formed by repeatedly selecting uniformly at random and *with replacement of m* examples from the original training set. This means that the same example may appear multiple times in the bootstrap replicate, or it may not appear at all. After completing all the rounds, a final combined classifier is formed which simply predicts with the majority vote of all of the base classifiers. The implementation procedure for bagging [194, 195] is given in table 8.1.

Table 8.1: Steps in Bagging ensemble method

Given the training data set = { (x_1,y_1), (x_2,y_2)...,(x_n,y_n)}

Suppose base learning algorithm = L, and number of learning rounds = T

1. For t=1,..., T

   1.1 Generate a boot strap sample from D as D_t = *boot strap (D)*

   1.2 Train a base learner h_t *from the boot strap sample,* $h_t = L(D_t)$

   End

2 Compute the output of the combined classifier as the class with the highest number of votes, $H(x)$ = majority $(h_1(x),...,h_t(x))$

8.4.2 Implementation using Boosting Ensemble Framework

The learning process of boosting method starts with uniform weighting. During each step of learning, weights of the examples which are not correctly learned by the weak learner are increased and those of the examples which are correctly learned by the weak learner are decreased. Strong classifiers are constructed by weighted voting of the weak classifiers. Since speech recognition is a multi-class pattern recognition problem, and boosting is applied only to binary classification problems, AdaBoost algorithm [194, 195] which allows multiple classes is used in this work. The AdaBoost algorithm is given in table 8.2.

Table 8.2: Steps in AdaBoost ensemble algorithm

Given the training data set = { (x_1,y_1), (x_2,y_2)...,(x_n,y_n)}

Suppose base learning algorithm = L, and number of learning rounds = T

1. Initialise the weight distribution $W_1(i) = 1/N$
2. For t=1,..., T

   2.1 $h_t = L(D,W_t)$

   2.2 e_t = *error of* h_t

   2.3 Determine the weight of h_t *as* $\alpha_t = \frac{1}{2} ln\ (1- e_t)/ e_t$

   2.4 Update the weight distribution $W_{t+1}(i) = \dfrac{w_t(i)\exp\left(-\alpha_t y_{(i)} h_t\left(x_{(i)}\right)\right)}{Z_t}$

where Z_t is a normalization factor

End

3 Compute the output of the combined classifier

$H(x) = \text{sign}(f(x)) = \text{sign}\ \sum_{t=1}^{T} \alpha_t h_t(x)$

8.4.3 Implementation using Stacking Ensemble Framework

Ensemble learning based on stacking combines multiple classifiers generated by using different learning algorithms on a single dataset. The dataset consists of pairs of feature vectors and their classifications. In the first phase, a set of base-level classifiers are generated and in the second phase, a meta-level classifier is learned that combines the outputs of the base-level classifiers. Table 8.3 shows the algorithm for stacking ensemble framework.

Table 8.3: Steps in Stacking ensemble method

Given the training data set D = { (x_1,y_1), (x_2,y_2)...,(x_n,y_n)}

Let the first level learning algorithms be L_1,....,L_T

Suppose second level learning algorithm = L

1. For t=1,..., T

   1.1. Train first level individual learner as $h_t = L_t(D)$

   End

2. Generate a new data set D' = ∅

3. For i = 1,..., N

   3.1. For t = 1,...,T

   3.1.1. $z_{it} = h_t(x_i)$

   End

   3.2. Combine the outputs of the base level classifiers,

   $D' = D' \cup \{((z_{i1},.....,z_{iT}),y_{(i)})\}$

   End

4. Train the second level learner h' = L(D')

5. Calculate the output as $H(x) = h'(h_1(x),....,h_T(x))$

8.5 Experimental Results using Ensemble Methods

So, here, the predicted classifications from MLP, SVM and Naive Bayes classifier can be used as input variables into a neural network meta-classifier. A comparison of the performance of the ensemble techniques bagging, boosting and stacking are given in table 8.4.

Table 8.4: Comparison of classification results

Technique used	Precision	Recall	Correctly Classified	Accuracy %
Bagging	0.983	0.983	19661	98.31
Boosting	0.946	0.942	18832	94.16
Stacking	0.962	0.96	19203	96.01

A graph showing the comparison of results obtained using ensemble classification methods namely, Bagging Boosting and Stacking are given in figure 8.4.

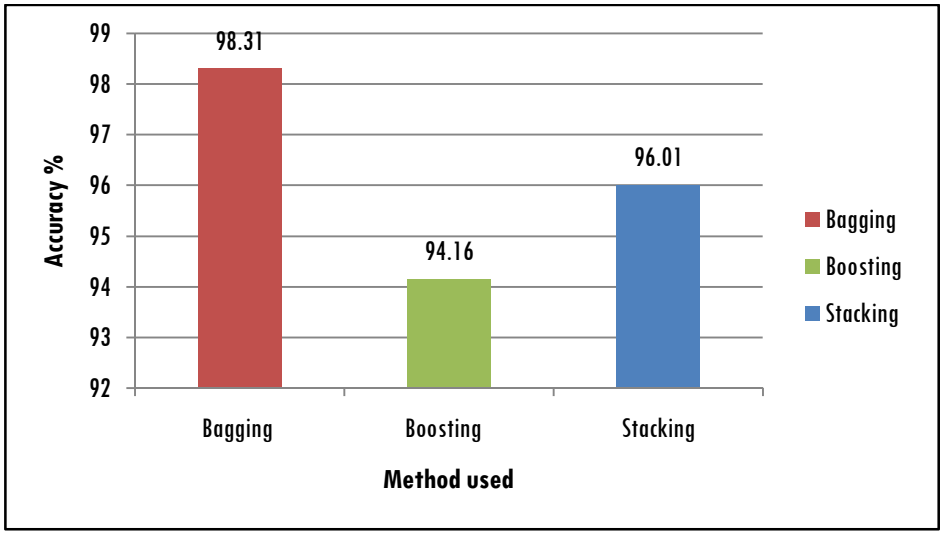


Figure 8.4: Comparison of results obtained using ensemble classification

Since better results are obtained using the ensemble classification method bagging, the confusion matrix obtained using this method is given in figure 8.5.

Predicted Class

Actual Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	990	0	1	0	0	1	0	0	0	1	0	1	0	0	2	2	1	1	0	0
2	0	992	0	0	0	5	2	0	0	0	0	0	0	1	0	0	0	0	0	0
3	0	3	966	0	0	3	6	7	0	2	0	1	1	2	0	4	0	0	0	5
4	1	3	5	915	13	1	20	1	9	0	3	6	3	0	2	2	3	3	6	4
5	0	0	2	1	990	0	1	1	0	0	1	1	0	0	0	1	0	0	2	0
6	0	0	0	0	1	993	3	1	0	0	0	0	0	0	1	0	0	1	0	0
7	0	3	0	2	3	0	976	1	3	0	0	2	1	4	0	5	0	0	0	0
8	0	1	0	1	1	0	0	990	0	1	1	1	0	1	3	0	0	0	0	0
9	0	2	0	0	0	1	1	0	989	0	0	1	0	0	0	1	2	1	1	1
10	1	1	0	0	0	0	0	5	0	989	0	1	0	1	1	1	0	0	0	0
11	1	0	1	2	1	0	0	0	0	0	991	1	0	0	1	1	0	0	1	0
12	0	2	0	0	4	2	7	0	1	2	1	967	0	8	2	4	0	0	0	0
13	1	0	2	0	0	0	0	1	2	1	1	0	990	0	0	0	0	0	2	0
14	0	1	1	0	0	0	1	0	0	0	0	3	0	991	2	1	0	0	0	0
15	0	0	1	1	0	0	0	2	0	1	1	1	0	2	990	0	0	0	1	0
16	0	0	2	0	0	0	1	3	0	0	0	0	0	0	2	990	0	1	1	0
17	3	0	0	1	2	0	0	0	1	0	0	1	0	0	0	0	990	1	1	0
18	0	0	0	2	2	0	0	1	0	0	0	1	0	0	1	1	1	990	1	0
19	0	1	0	0	2	0	0	0	1	1	1	2	1	0	0	0	0	0	991	0
20	0	1	0	0	2	0	0	3	2	4	0	1	0	2	1	1	0	0	2	981

Figure 8.5: Confusion matrix for ensemble classification using Bagging

8.6 Comparison of the Performance of Speech Recognition Systems Developed

This section is intended to provide a comparison of the various speech recognition systems developed as a part of this research work. The evaluation of the results is represented in terms of recognition accuracy. The table 8.5 given below shows a summary of all the results obtained from the experiments done for developing a speech recognition system with utmost accuracy.

Table 8.5: Performance evaluation of all the speech recognition systems developed using different pre-processing, feature extraction, post processing and classification techniques

Pre-processing method	Feature extraction Method	Post processing method		Classifier		Recognition Accuracy
End Point Detection, Pre- emphasis, Framing, Windowing	Linear Predictive Coding	Normalisation		Artificial Neural Networks		79.89
				Hidden Markov Models		75.5
				Naive Bayes		76.62
				Support Vector Machines		78.01
End Point detection, Pre-emphasis, Framing, Windowing	Mel Frequency Cepstral Coefficients	Normalisation		Artificial Neural Networks		81.77
				Hidden Markov Models		77.26
				Naive Bayes		78.48
				Support Vector Machines		80.27
End Point Detection, Soft Thresholding	Discrete Wavelet Transforms	Normalisation		Artificial Neural Networks		87.21
				Hidden Markov Models		83.01
				Naive Bayes		84.12
				Support Vector Machines		86.08
End Point Detection, Soft Thresholding	Wavelet Packet Decomposition	Normalisation		Artificial Neural Networks		86.84
				Hidden Markov Models		82.57
				Naive Bayes		83.76
				Support Vector Machines		85.53
End Point Detection, Soft Thresholding	Discrete Wavelet Packet Decomposition	Principal Component Analysis		Artificial Neural Networks		89.50
End Point Detection, Adaptive Smoothing, Soft Thresholding	Discrete Wavelet Packet Decomposition	Principal Component Analysis		Artificial Neural Networks		92.65
End Point Detection, Adaptive Smoothing Soft Thresholding	Discrete Wavelet Packet Decomposition	Principal Component Analysis	Three Sigma Limits	Artificial Neural Networks		93.77
			Quartile			95.78
			Mode			92.815
End Point Detection, Adaptive Smoothing Soft Thresholding	Discrete Wavelet Packet Decomposition	Principal Component Analysis, Quartile		Artificial Neural Networks, Support Vector Machines, Naive Bayes Classifier	Bagging	98.31
					Boosting	94.16
					Stacking	96.01

The comparison of the performance of all the speech recognition systems developed in this research work is given in the graph in figure 8.6.

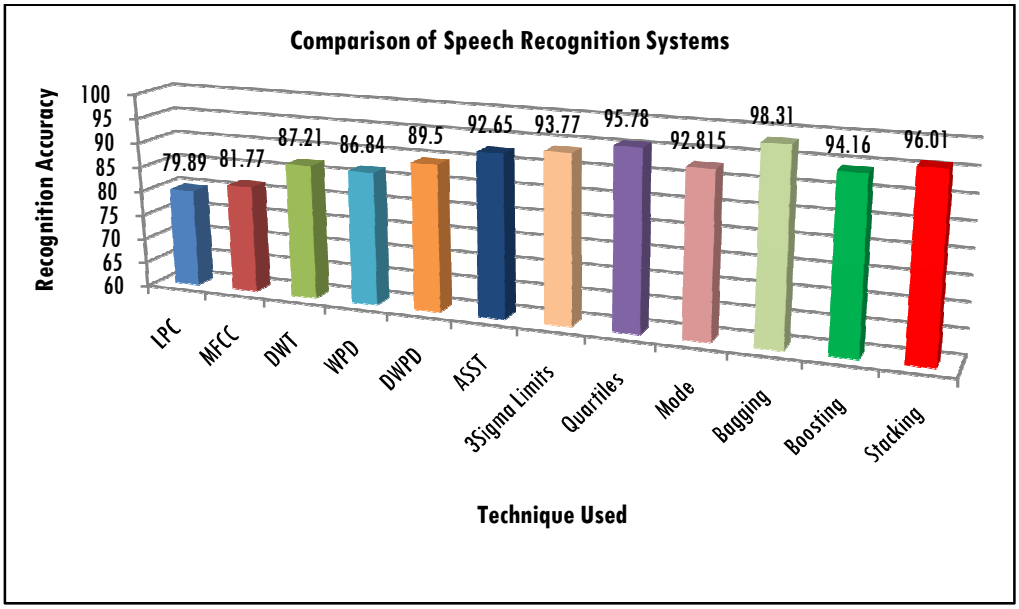


Figure 8.6: Comparison of different speech recognition systems developed

8.7 Summary of the Chapter

This chapter presents a framework for classification by combining different classifiers. Here we investigate whether an ensemble of classifiers can improve the recognition rate thereby improving the performance of the speech recognition system developed. For exploring this, both voting and stacking techniques are employed. This also analyses the performance of different ensemble learning methods on the speech database. Classifiers are combined using three ensemble learning techniques such as Bagging, Boosting and Stacking. The performance evaluation of these techniques in recognising the speech samples is carried out. Results obtained shows that an ensemble classifier technique is a good method of improving the accuracy of a group of classifiers by combining their results. An ensemble method which performed

well for a dataset may not be good for another one. The results obtained depend on the base classifiers and the technique used for combining. In this work, the best results were obtained using bagging ensemble learning.

This chapter also presents a summary of the performance of all the speech recognition systems developed in this work using different pre-processing, feature extraction, post processing and classification techniques. A comparison of all the results is performed. It is observed that the new algorithms and methods developed during each stage of development of the speech recognition system yield better results by improving the recognition rate.

Chapter 9

FUTURE DIRECTIONS AND CONCLUSION

Contents

This thesis primarily addresses the problem of automatic speech recognition for recognising isolated words in the Malayalam language. Since the performance of a speech recognition system relies on the pre-processing steps, feature extraction techniques adopted, post processing methods applied on the feature vector set obtained and the pattern classifiers used, the main objective of this work is to build a speech recognition system with maximum recognition accuracy. To achieve this goal, new algorithms and improvements are necessary at each stage of the speech recognition process. Different performance assessment measures are employed for evaluating the performance of the speech recognition systems developed. These representation models are proved to be effective in improving the recognition rate. So this research work proposes new enhanced algorithms and improved techniques for building an efficient speech recognition system.

9.1 Summary of the Research Work

In spite of the advances in technology and hardware during the last few decades, Automatic Speech Recognition is still a challenging and difficult task when it comes to real world applications. Now-a-days, speech recognition has wide applications in almost all fields of life. Due to this wide variety of applications, the requirements for each application are different. Researchers

are therefore trying to explore effective ways to build efficient speech recognition systems for each application. This work intends to enhance the performance of the already existing methods to improve the recognition rate.

From the literature survey conducted, it was not possible to rule out a specific technique, which always produces the best results. So in this research work, first a study is carried out to find a suitable combination of techniques for the efficient recognition of the speech samples in the databases created for the Malayalam language. The present work can be considered to have two phases. In the first phase, sixteen different speech recognition systems are developed to find the best feature extraction technique and pattern classifier combination which produced the best recognition rate for the speech databases created in Malayalam. During the second phase, new enhanced algorithms and improvements are proposed, designed and developed to further improve the recognition rate. The major highlights of both phases are given below.

The main highlights of the activities in the first phase of the research work are:

- Creation of three databases in Malayalam for vowels, digits and isolated words with 100 speakers, 200 speakers and 1000 speakers respectively.
- Exploitation of End Point Detection algorithm and Pre-emphasis Filters for noise reduction and speech enhancement.
- Development of a speech recognition system using Linear Predictive Coding (LPC) parameters and its implementation using four different classifiers like Artificial Neural Networks (ANN), Support Vector Machines (SVM), Hidden Markov Models (HMM) and Naive Bayes Classifiers.

- Evaluation of the consistency of the features based on recognition accuracy depending on the number of speakers.
- Implementation of Mel Frequency Cepstral Coefficients (MFCC) features for the recognition of Malayalam speech databases using different classifiers.
- Exploitation of wavelet denoising algorithms based on Soft Thresholding for denoising of signals.
- Experiments done for selecting the best wavelet family and mother wavelet for the databases created.
- Design and implementation of a speech recognition system for the three databases created using wavelet based Discrete Wavelet Transforms (DWT) feature extraction technique using various classifiers.
- Evaluating the performance of the speech recognition system developed using Wavelet Packet Decomposition (WPD) feature vector set and the above given classifiers.
- Comparison of the performance of these feature extraction techniques and classifiers to select the combination of feature extraction technique and classifier which performed best for Malayalam in terms of recognition accuracy.

The main highlights of the activities in the second phase of the research work are:

- Introduction of a new improved algorithm for feature extraction called Discrete Wavelet Packet Decomposition (DWPD) for the better performance of the speech recognition system. The characteristics of

different feature extraction techniques are combined to create a new hybrid algorithm which can produce better results based on recognition accuracy.

- Exploitation of Principal Component Analysis (PCA) technique for effective reduction of the dimension of the feature vector set without much affecting the recognition accuracy. PCA provides the dominant traits on which greater emphasis is laid.
- Devising of a new algorithm for smoothing the speech signals before pre-processing for reducing the amount of noise present in the signals. Noise is omnipotent and an algorithm that can weed out the sudden spikes which are only mere disturbances helps in enhancing the quality of the original signal. This proposed Adaptive Smoothing algorithm along with the wavelet denoising method based on Soft Thresholding called Adaptive Smoothing Soft Thresholding (ASST) helps in the enhancement of the signal by reducing the Signal-to-Noise Ratio and the error produced.
- Formulation of a new method for post processing based on the statistical technique of Three Sigma Control Limits which utilises the features of Mean. Statistical methods provide better uniformity among the data values in the feature vector set and also ensure reliable identification.
- Introduction of a new method for limiting the range of data using statistical technique based on Quartiles which uses the characteristics of Median. The performance of the feature vectors varies with the selection of the range. Choosing a range that can even out the variations in the data set helps in better analysis and classification.

- Devised a new method that can be applied during the post processing stage based on the statistical mode calculation.
- Investigation of ensemble classifiers based on three ensemble learning methods namely Bagging, Boosting and Stacking by combining the various classifiers for the better performance of the system. This hybrid architecture of combining different classifiers using different ensemble learning methods overcomes the limitations of using a single classifier. Ensemble learning utilises the best classifier combinations based on two popular schemes namely Voting and Stacking.
- Achievement of encouraging and improved recognition rates for all the four stages of development of a speech recognition system using the proposed new algorithms and methods applied during the pre-processing, feature extraction, post processing and classification stages.

9.2 Future Directions

In this research work, we have designed a speech recognition system with a fair degree of accuracy. Our main emphasis was on speech recognition for isolated words which finds applications in industry and man-machine interfaces. However, there is scope for further research in this field and some of the future prospects are listed below.

- **Extending the work to Continuous speech recognition:** This works concentrates only on the recognition of isolated words. The new proposed algorithms can also be tried on continuous speech since it also includes pre-processing, feature extraction, post processing and classification modules.

- **Expanding the work to large vocabulary systems:** This research work was carried out for only medium number of data such as 12 vowels, 10 digits and 20 isolated words. In future, this can be expanded to large vocabulary systems.
- **Extending the work to speaker recognition:** This work focuses only on recognising speech. Another area to which this research work can be extended is of speaker recognition. Speaker recognition deals with identifying the speaker instead of recognising what he says.
- **Extending the work to language independent speech recognition system:** This work is meant to design efficient speech recognition for Malayalam language. This work can also be extended to different languages since the architecture of the speech recognition system is the same.

9.3 Conclusion

Speech recognition is a complicated task and state-of-the-art recognition systems show that its performance depends on many factors like the number of speakers, the database used and the different techniques adopted during the different stages of development of the system. The main intention of this research work is to build a speech recognition system for recognising speaker independent isolated words in Malayalam with utmost recognition accuracy. So databases are created in Malayalam and experiments are performed therein. From the literature study, it was not possible to select a specific combination of the feature extraction method and a classifier which always generated good results. Hence sixteen different experiments were carried out for selecting the best combination with the best recognition rate using 4 feature extraction techniques and 4 pattern classifiers. Among these techniques, DWT and MLP combinations were found to produce the best

results. Among the different wavelet families available, better performance was obtained using the Daubechies family of wavelets with order 4 (db4).

In this research work, new algorithms and improvements were proposed, designed and developed during the four stages of development of the speech recognition system. The proposed hybrid algorithm DWPD which was developed during the feature extraction stage by combining the features of both DWT and WPD produced improvements in the degree of recognition accuracy. The newly proposed adaptive smoothing technique which was applied during the pre-processing stage played a significant role in removing the sudden spikes due to noise, thus improving the SNR value. These smoothed signals when applied to wavelet denoising using Soft Thresholding, yielded better recognition accuracy. All the three statistical thresholding techniques proposed during the post processing stage based on Three Sigma Limits, Quartiles and Confidence Interval Mode were found to be efficient in selecting the feature vectors for pattern recognition. These techniques were employed to bring the feature vectors to a particular predefined range. Among these three methods, the results obtained using Quartiles were proved to be superior. It was observed that the ensemble learning methods based on Bagging, Boosting and Stacking which were applied during the classification stage also generated better results. Thus the newly proposed algorithms and improved techniques performed well and produced better results for the speech recognition system developed for Malayalam.

The thesis findings can be used for different purposes with variations in language, databases etc. The consequent applications derived from the thesis findings will be on the rise.

The thesis findings can be used to develop similar applications in foreign languages and it can be used in Automated Teller Machines (ATMs) for dispensing cash and other financial transactions across all languages. This will allow foreigners as well as citizens of the country, a large degree of freedom from language-barriers in conducting financial transactions. The spoken digits recognition system has great relevance in this field.

An efficient speaker independent isolated words recognition system has a number of applications in different fields. It has applications across the Internet in helping farmers and other less literate people to access international markets for information on commodities and future pricing positions. It has also a wide range of applications in robotics where actions and tasks can be executed/cancelled using voice commands in native languages. The need for high- level technical expertise in accessing state- of- the- art technologies can be transgressed and brought down within the reach of the common man. The automotive industry which uses robots can enhance their productivity by including voice detection. Strategically placed in airports and other places of public interest, it can be used to locate criminals whose voice data is already available. It can help investigating agencies in tracking down criminals. It can be used to supplement security measures by bringing in additional check measures to establish authenticity. The spoken words recognition system is of great relevance in this context.

The findings can also be used in our quest to derive unspoken words from an existing dataset. A database of consonants and vowels can be developed and maintained to create speech signals by combining the consonants and vowels. This can mimic an undelivered speech and identify potential speakers for a particular theme or event. A database of all consonants

and vowels can create a dictionary of all words that can be spoken by an individual. This can be compared with actual spoken words to verify identity.

Presently all dictionaries are language specific. English-English, English – Malayalam etc. With a voice recognition system in place, we can develop dictionaries that can transcend language barriers. Suppose we have an English – Malayalam dictionary and Malayalam – Hindi dictionary, an English – Hindi dictionary can be only a few seconds away in the hands of a software personnel. In a similar way we can develop dictionaries against all languages that will ultimately remove barriers of language. The works of speech translators can also be made eas**y.** In the recent past we have seen translators working hard during visit by foreign dignitaries. A speaker and a hearing mechanism can covert alien language to ones own mother tongue without waiting for a translator.

The music industry also finds varied uses. The ragas developed in ancient ages are prone to dilution in the hands of inexperienced artists. Music competitions are won based on the judgements by the judges who may or may not be right. Presently the ragas and other intricate music systems can be verified using voice recognition systems. The feature vectors corresponding to the original raga can be stored and compared with the performances by artists.

Finally, this research work has been a comprehensive approach for the development of a speech recognition system with emphasis on all the different aspects namely, the pre-processing steps, feature extraction techniques, post processing methods and the classification techniques. No one technique is perfect in itself and we have therefore adopted a hybrid architectural approach.

REFERENCES

[1] Lawrence R. Rabiner, and Ronald W. Schafer, "Introduction to Digital Speech Processing," *Foundations and Trends in Signal Processing*, vol. 1, nos. 1–2, pp. 1-194, Jan. 2007.

[2] Samudravijaya K., "Speech and Speaker recognition: a tutorial," in *Proc. International Workshop on Technology Development in Indian Languages*, Kolkata, Jan. 2003.

[3] Kenneth Thomas Schutte, "Parts-based Models and Local Features for Automatic Speech recognition," Ph.D. dissertation, Dept. of Elec. Eng. and Comp. Sci., Massachusetts Inst. of Tech., Massachusetts, 2009.

[4] O. Scharenborg, "Reaching Over the Gap: A Review of Efforts to Link Human and Automatic Speech Recognition Research," *Speech Communication*, vol. 49, pp. 336–347, May 2007.

[5] R. P. Lippmann, "Speech Recognition by Machines and Humans," *Speech Communication*, vol. 22, pp. 1–15, Apr. 1997.

[6] Kuldeep Kumar, and R. K. Aggarwal, "Hindi Speech Recognition System Using Htk," *International Journal of Computing and Business Research*, vol. 2, no. 2, pp. 2229-6166, May 2011.

[7] George L. Hart. (2000, Apr.). Statement on the Status of Tamil as a Classical Language, Letter on Tamil as a Classical Language. Univ. of California, Berkeley. [Online]. Available: http://tamil.berkeley.edu/tamil-chair/letter-on-tamil-as-a-classical-language

[8] M. A. Anusuya, and S. K. Katti, "Speech Recognition by Machine: A Review," *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 6, no. 3, pp. 181-205, Dec. 2009.

[9] Xiao Xiong, "Robust Speech Features and Acoustic Models for Speech Recognition," Ph.D. dissertation, School of Comp. Eng., Nanyang Technological Univ., Singapore, 2009.

[10] Markus Forsberg. (2003). *Why is Speech Recognition Difficult*? [Online]. Available: www.speech.kth.se/~rolf/gslt_papers/MarkusForsberg.pdf

[11] David Gerhard, "Audio Signal Classification: History and Current Techniques," Dept. of Comp. Sci., Univ. of Regina, Regina, Rep. TR-CS 2003-07, 2003.

[12] L. R. Rabiner, and B. H. Juang, *Fundamentals of Speech Recognition*, New Jersey, USA: Engle-wood Cliffs Publisher, 1993.

[13] Jeremy Bradbury. (2000). *Linear Predictive Coding* [Online]. Available: http://my.fit.edu/~vKepuska/ece5525/lpc_paper.pdf

[14] Thomas F. Quatieri, *Discrete- Time Speech Signal Processing principles and Practice*, New Jersey, USA: Pearson Education Inc., 2002.

[15] B. Plannerer. (2005). *An Introduction to Speech Recognition* [Online]. Available: http://www.speech-recognition.de/pdf/introSR.pdf

[16] Hemdal J. F., and G. W. Hughes, "A feature based computer recognition program for the modeling of vowel perception," in *W. Wathen-Dunn Ed. Models for the Perception of Speech and Visual Form*, Massachusetts, USA: MIT Press, 1967.

[17] F. Itakura, "Minimum Prediction Residual Applied to Speech Recognition," *IEEE Trans. Acoustics, Speech, Signal Proc*., vol. 23, no. 1, pp. 67-72, Feb. 1975.

[18] Santhosh K. Gaikwad, Bharti W. Gawali, and Pravin Yannawar, "A Review on Speech Recognition Technique", *International Journal of Computer Applications,* vol. 10, no. 3, pp. 16-24, Nov. 2010.

[19] R.K.Moore, "Twenty things we still don't know about speech," in *Proc. of CRIM/ FORWISS Workshop on Progress and Prospects of speech Research and Technology*, Germany, Jul. 1994.

[20] Shanthi Therese S., and Chelpa Lingam, "Review of Feature Extraction Techniques in Automatic Speech Recognition," *International Journal of Scientific Engineering and Technology*, vol. 2, no. 6, pp. 479-484, Jun. 2013.

[21] Vimala C., and Dr. V. Radha, "A Review on Speech Recognition Challenges and Approaches," *World of Computer Science and Information Technology Journal (WCSIT)*, vol. 2, no. 1, pp. 1-7, Jan. 2012.

[22] Lawrence R. Rabiner, "Applications of Speech Recognition in the Area of Telecommunications," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 501-510, Santa Barbara, Dec. 1997.

[23] B.H. Juang, and L. R. Rabiner, "Automatic Speech Recognition – A Brief History of the Technology Development," *Elsevier Encyclopedia of Language and Linguistics*, pp. 1-24, Aug. 2004.

[24] Sadaoki Furui, "50 Years of Progress in Speech and Speaker Recognition Research," *ECTI Transactions On Computer And Information Technology*, vol. 1, no. 2, pp. 64-74, Nov. 2005.

[25] Wiqas Ghai, and Navdeep Singh, "Literature Review on Automatic Speech Recognition," *International Journal of Computer Applications*, vol. 41, no. 8, pp. 42-50, Mar. 2012.

[26] Biing Hwang Juang, and Sadaoki Furui, "Automatic Recognition and Understanding of Spoken Language—A First Step Toward Natural Human–Machine Communication," *Proc. of the IEEE*, vol. 88, No. 8, pp. 1142-1165, Aug. 2000.

[27] Sanjivani S. Bhabad, and Gajanan K. Kharate, "An Overview of Technical Progress in Speech Recognition," *International Journal of*

Advanced Research in Computer Science and Software Engineering, vol. 3, no. 3, pp. 488-497, Mar. 2013.

[28] Preeti Saini, and Parneet Kaur, "Automatic Speech Recognition: A Review," *International Journal of Engineering Trends and Technology*, vol. 4, no. 2, pp. 132-136, 2013

[29] Md Salam, Dzulkifli Mohamad, and Sheikh Salleh, "Malay Isolated Speech Recognition Using Neural Network: A Work in Finding Number of Hidden Nodes and Learning Parameters," *The International Arab Journal of Information Technology*, vol. 8, no. 4, pp. 364-371, Oct. 2011.

[30] Thiang, and Suryo Wijoyo, "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot," in *Proc. International Conference on Information and Electronics Engineering (IPCSIT)*, vol. 6, Bangkok, May 2011, pp. 179-183.

[31] Sonia Sunny, David Peter S., and K. Poulose Jacob, "A Comparative Study of Parametric Coding and Wavelet Coding Based Feature Extraction Techniques in Recognizing Spoken Words," in *Proc. CUBE International Information Technology Conference and Exhibition*, Pune, Sept. 2012, pp. 326- 331.

[32] Preeti Saini, Parneet Kaur, and Mohit Dua, "Hindi Automatic Speech Recognition Using HTK," *International Journal of Engineering Trends and Technology (IJETT*), vol. 4, no. 6, pp. 2223-2229, Jun. 2013.

[33] Noraini Seman, Zainab Abu Bakar, and Nordin Abu Bakar, "Measuring the Performance of Isolated Spoken Malay Speech Recognition Using Multi-layer Neural Networks," in *Proc. of International Conference on Science and Social Research (CSSR 2010)*, Malaysia, Dec. 2010, pp. 182-186.

[34] Omesh Wadhwani, Amit Kolhe, and Sanjay Dekate, "Recognition of Vernacular Language Speech for Discrete Words using Linear Predictive

Coding Technique," *International Journal of Soft Computing and Engineering (IJSCE)* , vol. 1, no. 5, pp. 188-192, Nov. 2011.

[35] J. R. Karam, W. J. Phillips, and W. Robertson, "New Low Rate Wavelet Models for the Recognition of Single Spoken Digits," in *Proc. Canadian Conference on Electrical and Computer Engineering*, vol. 1, pp. 331 – 334, Mar. 2000.

[36] K. Daqrouq, A. R. Al-Qawasmi, K.Y. Al Azzawi, and T. Abu Hilal, "Discrete Wavelet Transform & Linear Prediction Coding Based Method for Speech Recognition via Neural Network, Discrete Wavelet Transforms - Biomedical Applications," in *Prof. Hannu Olkkonen Ed. InTech*, Sep. 2011.

[37] Roopa A. Thorat, and Ruchira A. Jadhav, "Speech Recognition System," in *Proc. International Conference on Advances in Computing, Communication and Control (ICAC3'09)*, New York, 2009, pp. 607- 609.

[38] Dr.Yousra F., and Al-Irhaim Enaam Ghanem Saeed, "Arabic Word Recognition Using Wavelet Neural Network," in *Proc of Third Scientific Conference in Information Technology*, Nov. 2010, pp. 416-425.

[39] Vimal Krishnan V. R., and Babu Anto P., "Features of Wavelet Packet Decomposition and Discrete Wavelet Transform for Malayalam Speech Recognition," *International Journal of Recent Trends in Engineering*, vol. 1, no. 2, pp. 93-96, May 2009.

[40] T. M. Thasleema, N. K. Narayanan, and N. S. Sreekanth, "A Robust Approach for Malayalam CV Speech Unit Recognition using Artificial Neural Network," in *Proc. International Conference on Image Processing, Computer Vision and Pattern Recognition*, vol. 1, 2011, pp. 52-56.

[41] Sherin M. Youssef, "A Robust Automated Speech Classification Using Hybrid Wavelet-based Architecture," in *Proc. 25th National Radio Science Conference* , Tanta, Mar. 2008, pp. 1-8.

[42] Yousef Ajami Alotaibi, "Comparative Study of ANN and HMM to Arabic Digits Recognition Systems," *Proc. JKAU: Eng. Sci*., vol. 19, no. 1, pp. 43-60, 2008.

[43] Md. Ali Hossain, Md. Mijanur Rahman, Uzzal Kumar Prodhan, and Md. Farukuzzaman Khan, "Implementation of Back-Propagation Neural Network For Isolated Bangla Speech Recognition," *International Journal of Information Sciences and Techniques (IJIST)*, vol. 3, no. 4, pp. 1-9, Jul. 2013.

[44] Gajanan Pandurang Khetri, Satish L. Padme, Dinesh Chnadra Jain, Dr. H. S. Fadewar, Dr. B.R. Sontakke, and Dr. Vrushsen P. Pawar, "Human Computer Interpreting with Biometric Recognition System," *International Journal of Application or Innovation in Engineering & Management (IJAIEM),* vol. 1, no.3, pp. 69-74, Nov. 2012.

[45] Ramón Fernández-Lorenzana, Fernando Pérez-Cruz, José Miguel García-Cabellos, Carmen Peláez-Moreno, Ascensión Gallardo-Antolín, and Fernando Díaz-de-María, "Some Experiments on Speaker-Independent Isolated Digit Recognition using SVM classifiers," *ITRW on Non-Linear Speech Processing*, Le Croisic, May 2003.

[46] Muhammad G., Alotaibi Y. A., and Huda M. N., "Automatic speech recognition for Bangla digits," in *Proc. 12th International Conference on Computers and Information Technology*, Dhaka, Dec. 2009, pp. 379-383.

[47] Antanas Lipeika, Joana Lipeikien E., and Laimutis Telksnys, "Development of Isolated Word Speech Recognition System," *INFORMATICA*, vol. 13, no. 1, pp. 37-46, 2002.

[48] Matthew K. Luka, Ibikunle A. Frank, and Gregory Onwodi, "Neural Network Based Hausa Language Speech Recognition," *International Journal of Advanced Research in Artificial Intelligence (IJARAI)*, vol. 1, no. 2, pp. 39-44, May 2012.

[49] N. S. Nehe, and R. S. Holambe, "DWT and LPC based Feature Extraction Methods for Isolated Word Recognition," *EURASIP Journal on Audio, Speech and Music Processing (Springer Open Journal)*, pp. 1-7, Jan. 2012.

[50] Shady Y. El-Mashed, Mohammed I. Sharway, and Hala H. Zayed, "Speaker Independent Arabic Speech Recognition Using Support Vector Machine," pp. 401-416. http://www.ektf.hu/agriamedia/data/present/209/209_present.pdf

[51] Sreejith C., and Reghuraj P. C., "Isolated Spoken Word Identification in Malayalam using Mel-frequency Cepstral Coefficients and K-means clustering," *International Journal of Science and Research (IJSR)*, vol. 1, no. 3, pp. 163-167, Dec. 2012.

[52] Mansour M. Alghamdi, and Yousef Ajami Alotaibi, "HMM Automatic Speech Recognition System of Arabic Alphadigits," *Arabian Journal for Science and Engineering*, vol. 35, no. 2C, pp. 137-155, Dec. 2010.

[53] Bassam A. Q. Al-Qatab, and Raja N. Ainon, "Arabic Speech Recognition Using Hidden Markov Model Toolkit(HTK)," in *Proc. International Symposium in Information Technology (ITSim)*, vol. 2, Kuala Lumpur, Jun. 2010, pp. 557-562.

[54] Ling He, Margaret Lech, Namunu C. Maddage, and Nicholas Allen, "Neural Networks and TEO features for an Automatic Recognition of Stress in Spontaneous Speech," in *Proc. Fifth International Conference on Natural Computation*, Tianjin, Aug. 2009, pp. 221- 231.

[55] Bishnu Prasad Das, and Rajan Parekh, "Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers," *International Journal of Modern Engineering Research (IJMER),* vol. 2, issue 3, pp. 854-858, May-Jun. 2012.

[56] Meysam Mohamad pour, and Fardad Farokhi, "A new approach for Persian speech Recognition," in *Proc. IEEE International Advance Computing Conference (IACC 2009)*, Patiala, Mar. 2009, pp. 153-158.

[57] Hemakumar G., and Punitha P., "Speaker Independent Isolated Kannada Word Recognizer," *Multimedia Processing, Communication and Computing Applications, Lecture Notes in Electrical Engineering*, vol. 213, pp. 333-345, Dec. 2012.

[58] Sukumar A. R., Shah A. F., and Anto P. B., "Isolated question words recognition from speech queries by using Artificial Neural Networks," in *Proc. International Conference on Computing Communication and Networking Technologies (ICCCNT)* , Karur, Jul. 2010, pp. 1-4.

[59] N. Uma Maheswari, A. P. Kabilan, and R. Venkatesh, "A Hybrid model of Neural Network Approach for Speaker independent Word Recognition," *International Journal of Computer Theory and Engineering*, vol. 2, no. 6, pp. 912-915, Dec. 2010.

[60] Chadawan Ittichaichareon, Siwat Suksri, and Thaweesak Yingthawornsuk, "Speech Recognition using MFCC," in *Proc. International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012)*, Pattaya, Jul. 2012, pp. 135-138.

[61] Javed Ashraf, Naveed Iqbal, Naveed Sarfraz Khattak, and Ather Mohsin Zaidi**, "**Speaker independent Urdu speech recognition using HMM,**"** in *Proc. of the 7th International Conference on Informatics and Systems (INFOS),* Cairo, Mar. 2010, pp.1-5.

[62] Cini Kurian, and Kannan Balakrishnan, "Malayalam Isolated Digit Recognition Using HMM and PLP Cepstral Coefficient," *International Journal of Advanced Information Technology (IJAIT),* vol. 1, no. 5, pp. 31-38, Oct. 2011.

[63] Engin Avci and Zuhtu Hakan Akpolat, "Speech recognition using a wavelet packet adaptive network based fuzzy inference system," *Expert Systems with Applications,* vol. 31, no. 3, pp. 495–503, Oct. 2006.

[64] Md. Akkas Ali, Manwar Hossain and Mohammad Nuruzzaman Bhuiyan, "Automatic Speech Recognition Technique for Bangla Words," *International Journal of Advanced Science and Technology,* vol. 50, pp. 51-60, Jan. 2013.

[65] Malay Kumar, R. K. Aggarwal, Gaurav Leekha, and Yogesh Kumar, "Ensemble Feature Extraction Modules for Improved Hindi Speech Recognition System," *IJCSI International Journal of Computer Science Issues*, vol. 9, issue 3, no. 1, pp. 175-181, May 2012.

[66] Leena R. Mehta, S. P. Mahajan, and Amol S. Dabhade, "Comparative Study of MFCC and LPC for Marathi Isolated Word Recognition System," *International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering,* vol. 2, no. 6, pp. 2133-2139, Jun. 2013.

[67] Mohit Dua, R. K. Aggarwal, Virender Kadyan, and Shelza Dua, "Punjabi Automatic Speech Recognition Using HTK," *International Journal of Computer Science Issues*, vol. 9, issue. 4, no. 1, pp. 359- 364, Jul. 2012.

[68] Shweta Bansal, "Isolated Word Speech Recognition System for Object Identification," *International Journal of Scientific & Engineering Research*, vol. 4, no.1, Mar. 2013.

[69] A. Akila, and E. Chandra, "Isolated Tamil Word Speech Recognition System Using HTK," *International Journal of Computer Science Research and Application,* vol. 3, Issue. 2, pp. 30-38, 2013.

[70] Baker J., Li Deng, Glass J., Khudanpur S., Chin-hui Lee, Morgan N., and O'Shaughnessy D., "Developments and directions in speech recognition and understanding, Part 1 [DSP Education]," IEEE Signal Processing Magazine, vol. 26, no. 3, pp. 75-80, May 2009.

[71] R. K. Aggarwal, and Mayank Dave, "Implementing a Speech Recognition System Interface for Indian Languages," in *Proc. IJCNLP-08 Workshop on NLP for Less Privileged Languages*, Hyderabad, Jan. 2008, pp. 105-112.

[72] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, issue 10-11, pp. 763–786, Nov. 2007.

[73] http://en.wikipedia.org/wiki/GoldWave

[74] James H. Martin, and Daniel Jurafsky, *Speech and Language Processing, An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed. New Jersey, USA: Prentice Hall, 2008.

[75] Jhing Fa Wang, and Shi-Huang Chen, "Wavelet Transforms for Speech signal processing," *Journal of the Chinese Institute of engineers*, vol. 22, no. 5, pp. 549-560, 1999.

[76] Bhupinder Singh, Rupinder Kaur, Nidhi Devgun, and Ramandeep Kaur, "The process of Feature Extraction in Automatic Speech Recognition System for Computer Machine Interaction with Humans," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, issue 2, Feb. 2012.

[77] A. B. M. Shawkat Ali, and Saleh A. Wasimi, *Data Mining: Methods and Techniques*, 3rd ed. New Delhi, India: Cengage Learning, 2009.

[78] H. Liu, and H. Motoda, *Feature Selection for Knowledge Discovery and data Mining*, Boston, Massachusetts: Kluwer Academic Publication, 1998.

[79] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, 2nd ed. U.K.: John Wiely & Sons Inc., 2000.

[80] Neeta Awasthy, J. P. Saini, and D. S. Chauhan, "Spectral Analysis of Speech: A New Technique," *International Journal of Information and Communication Engineering*, vol. 2, no. 1, pp. 19-28, Jan. 2006.

[81] Qi Li, Jinsong Zheng, Augustine Tsai, and Qiru Zhou, "Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition," *IEEE Transactions on Speech and Audio Processing,* vol. 10, no. 3, pp. 146-157, Mar. 2002.

[82] Kapil Sharma, H. P. Sinha, and R. K. Aggarwal, "Comparative Study of Speech Recognition System Using Various Feature Extraction Techniques," *International Journal of Information Technology and Knowledge Management,* vol. 3, no. 2, pp. 695-698, Jul.-Dec. 2010.

[83] Atal B., and Rabiner L., "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, issue 3, pp. 201 – 212, Jan. 2003.

[84] D. G. Childers, M. Hand, and J. M. Larar, "Silent and Voiced/Unvoied/ Mixed Excitation (Four-Way), Classification of Speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1771-1774, Nov. 1989.

[85] Dongzhi He, Yibin Hou, Yuanyuan Li, and Zhi-Hao Ding, "Key Technologies of Pre-processing and Post-processing methods for Embedded Automatic Speech Recognition Systems," in *Proc. of IEEE/ASME International Conference on Mechatronics and Embedded Systems and Applications (MESA),* Qingdao, Jul. 2010, pp. 76-80.

[86] Li Deng, and D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach*, 1st ed. USA: CRC Press, 2003.

[87] Sankar K. Pal and Pabitra Mitra, *Pattern recognition algorithms for Data Mining*, 1st ed. Boca Raton London, New York: Chapman and Hall/CRC, 2004.

[88] Brian D. Ripley, *Pattern Recognition and Neural Networks*, 1st ed. Cambridge, New York: Cambridge University Press, 2008.

[89] Christopher M. Bishop, *Neural Networks for Pattern Recognition*, 1st ed. Oxford University Press, 1996.

[90] S. N. Sivanadam, S. Sumathi, and S. N. Deepa, *Introduction to Neural Networks using Matlab 6.0*, New Delhi, India: Tata McGraw-Hill, 2006.

[91] James A. Freeman, and David M. Skapura, *Neural Networks Algorithm, Application and Programming Techniques*, California, USA: Pearson Education, 2006.

[92] Wouter Gevaert, Georgi Tsenov, and Valeri Mladenov, "Neural Networks used for Speech Recognition," *Journal of Automatic Control*, vol. 20, pp. 1-7, 2010.

[93] Richard P. Lippmann, "Neural Network Classifiers for Speech Recognition," *the Lincoln Laboratory Journal*, vol. 1, no. 1, pp. 107-124, 1988.

[94] Laurene Fausett, *Fundamentals of Neural Networks Architectures, Algorithms and Applications*, 1st ed. Prentice Hall, 1994.

[95] Ajith Abraham, "Artificial Neural Networks," in *Handbook of Measuring System Design*, vol. 1, Wiley, 2005, ch. 129, pp. 901-908.

[96] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revisited: A stepwise procedure for building and training a neural network," *Neurocomputing: Algorithms, Architectures and Applications*, vol. F 68 of NATO ASI Series, Springer - Verlag, 1990.

[97] V. N. Vapnik, *Statistical Learning Theory*, New York, USA: J. Wiley, 1998.

[98] N. Cristianini, and J. Shawe-Taylor, *An introduction to Support Vector Machines*, Cambridge, UK: Cambridge University Press, 2000.

[99] B. Scholkopf, and A. Smola, *Learning with Kernels*, Cambridge, UK: MIT Press, 2002.

[100] Christopher J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167 (1998) Kluwer Academic Publishers, pp. 1-47, 1998.

[101] Ulrich H.-G. Kreßel, "Pairwise Classification and Support Vector Machines," Advances in Kernel Methods Support Vector Machine Learning, Cambridge, MA, MIT press, pp. 255-268, 1999.

[102] C. W. Hsu, and C. J. Lin, "A Comparison of Methods for Multi-class Support Vector Machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[103] Li Dan, Liu Lihua, and Zhang Zhaoxin, "Research of Text Categorization on WEKA," in *Proc. Third International Conference on Intelligent System Design and Engineering Applications*, Hong Kong, Jan. 2013, pp. 1129-1131.

[104] J. Han, and M. Kamber, *Data Mining Concepts and Techniques*, 3rd ed. San Francisco, USA: Morgan Kaufmann Publishers, 2007.

[105] Laszlo Toth, Andras Kocsor, and Janos Csirik, "On Naive Bayes in Speech Recognition," *International Journal of Applied Mathematics and Computer Science*, vol. 15, no. 2, pp. 287–294, Jun. 2005.

[106] L. R. Rabiner, "A Tutorial on Hidden Markov Model and Selected Applications in Speech Recognition," *Proc. of the IEEE*, vol. 77, no. 2, pp. 257-286, Feb. 1989.

[107] B. H. Juang, L. R. Rabiner, "Hidden Markov Models for Speech Recognition," *Technometrics*, vol. 33, no. 3, pp. 251-272, Aug. 1991.

[108] Felinec, *Statistical methods for speech recognition*, Cambridge, USA: MIT press, 1997.

[109] Holmes J., and Holmes W., *Speech Synthesis and Recognition*, 2nd ed. London, U. K.: Tailor & Francis, 2001.

[110] Stephen J. Chapman, *MATLAB Programming for Engineers*, 4th ed. Canada: Thomson Learning, 2008.

[111] Amos Gilat, *MATLAB: An Introduction with Applications*, 1st ed. India: Wiley India, 2007.

[112] Baharak Goli, and Geetha Govindan, "Weka - A powerful free software for implementing Bio-inspired Algorithms," *Technical Trends, CSI Communications*, pp. 9-12, Dec. 2011.

[113] Nusharani, and P. N. Girija, "Error analysis to improve the speech recognition accuracyon Telugu language," *Sadhana*, vol. 37, part 6, Dec. 2012, pp. 747–761.

[114] Robi Polikar. (1999) *the Engineer's Ultimate Guide to Wavelet Analysis Wavelet Tutorial* [Online]. Available: http://users.rowan.edu/~polikar/WAVELETS/WTtutorial.html

[115] Daubechies J., *Ten Lectures on Wavelets*, 2nd ed. Philadelphia, USA: SIAM, 1992.

[116] Y. Meyer, *Ondelettes et Opérateurs, I: Ondellettes, II: Opérateurs de Calderón-Zygmund, III: (with R. Coifman), Opérateurs Multilinéaires*. Hermann, France: Cambridge Univ. Press, English translation of vol. 1, 1993.

[117] K.P. Soman, K.I. Ramachandran, and N.G. Resmi, *Insight into Wavelets from theory to practice*, 3rd ed. New Delhi, India: PHI learning private Limited, 2010.

[118] Dr. Shaila D. Apte, *Speech and audio processing*, India: Wiley India Pvt. Ltd., 2012.

[119] Walid G. Morsi, and M. E. El-Hawary, "The Most Suitable Mother Wavelet for Steady-State Power System Distorted Waveforms," in *Proc. IEEE Canadian conference on Electrical and Computer Engineering*, Niagara Falls, May 2008, pp. 17-22.

[120] Robi Polikar, "The story of wavelets," in *Proc. IMACS/IEEE CSCC*, 1999, pp. 5481-5486.

[121] Michael weeks, *Digital signal processing using MATLAB and wavelets*, 2nd ed. New Delhi, India: Firewall media, 2007.

[122] Y. Meyer, *Wavelets and Operators*, Cambridge, U. K.: Cambridge Univ. Press, 1992.

[123] S. Mallat, *A Wavelet Tour of Signal Processing*, 3rd ed. New York, USA: Academic, 1998.

[124] Fatma H. Elfouly, Mohamed I. Mahmoud, Moawad I. M. Dessouky, and Salah Deyab, "Comparison between Haar and Daubechies Wavelet Transformations on FPGA Technology," *International Journal of Computer and Information Engineering*, vol. 2, no. 1, pp. 37-42, 2008.

[125] DongMei Chen, "Multiresolution Models on Image Analysis and Classification," in *Proc. UCGIS Annual Assembly and Summer Retreat*, Maine, Jun. 1997.

[126] George Tzanetakis, Georg Essl, and Perry Cook, "Audio Analysis using the Discrete Wavelet Transform," in *Proc. WSES International Conference, Acoustics and Music: Theory and Applications*, 2001, pp. 318-323.

[127] S. Kadambe, and P. Srinivasan. "Application of adaptive wavelets for speech," *Optical Engineering*, vol. 33, no. 7, pp. 2204-2211, Oct. 1994.

[128] Mohammed Bahoura, and Jean Rouat, "Wavelet speech enhancement based on time-scale adaptation," *Speech Communication*, vol. 48, issue 12, pp. 1620–1637, Dec. 2006.

[129] Farooq, S. Datta, "Wavelet-based denoising for robust feature extraction for speech recognition," *Electronics Letters*, vol. 3, no. 1, pp. 163-165, 2003.

[130] D. L. Donoho, "De-noising by soft thresholding," *IEEE transactions on information theory*, vol. 41, no. 3, pp. 613-627, May 1995.

[131] Yasser Ghanbari, and Mohammad Reza Karami, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets," *Speech Communication*, vol. 48, no. 8, pp. 927–940, Aug. 2006.

[132] Tie Cai, and Xing Wu, "Wavelet-Based De-Noising of Speech Using Adaptive Decomposition," in *Proc. IEEE International Conference On Industrial Technology*, Chengdu, Apr. 2008, pp. 1-5.

[133] R. K. Martinet, J. Morlet, and A. Grossman, "Analysis of Sound Patterns through Wavelet Transform," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 1, no. 2, pp. 237-301, Aug. 1987.

[134] M. Vetterli, and C. Herley, "Wavelets and Filter Banks: Theory and Design," *IEEE Transactions on Signal Processing*, vol. 40, issue 9, pp. 2207- 2232, Sep. 1992.

[135] Gajanan K. Kharate, Varsha H. Patil, and Niranjan L. Bhale, "Selection of Mother Wavelet for Image Compression on Basis of Nature of Image," *Journal of Multimedia*, vol. 2, no. 6, pp. 44-51, Nov. 2007.

[136] "*Special Issue on Wavelets and Signal Processing*", *IEEE Trans. Signal Processing*, vol.41, Dec. 1993.

[137] S .G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, issue 7, pp. 674-693, Jul. 1989.

[138] Ting W., Guo-zheng Y., Banghua Y., and Hong S., "EEG Feature Extraction based on Wavelet Packet Decomposition for Brain Computer Interface," *Measurement*, vol. 41, no. 6 , pp. 618-625, Jul. 2008.

[139] Chan Woo S., Peng Lin C., and Osman R., "Development of a Speaker Recognition System using Wavelets and Artificial Neural Networks," in *Proc. International Symposium on Intelligent Multimedia, Video and Speech processing*, Hong Kong, May 2001, pp. 413-416.

[140] Fecit Science and Technology Production Research Center, "Wavelet Analysis and Application by MATLAB6.5 [M]", Electronics Industrial Press, Beijing, 2003.

[141] Ronald R. Coifman, and Mladen V. Wickerhauser, "Entropy based Algorithm for best Basis Selection," *IEEE Transactions on Information Theory*, vo1. 38, no. 2, pp. 713-718, Mar. 1992.

[142] B. C. Li, and J. S. Luo, *Wavelet Analysis and Its Applications*, Electronics Engineering Press, Beijing, China, 2003.

[143] Zhen-li Wang, Jie Yang, and Xiong-wei Zhang, "Combined Discrete Wavelet Transform and Wavelet Packet Decomposition for Speech Enhancement," in *Proc. 8th International Conference on Signal Processing*, vol. 2, Beijing, 2006.

[144] A. J. Richard, *Applied Multivariate Statistical Analysis*, 3rd ed. New Jersey, USA: Prentice hall, 1992.

[145] Xuechuan Wang, "Feature extraction and dimentianality reduction in pattern recognition and their application in speech recognition," Ph.D. dissertation, School of Microelectronical Engin., Griffith Univ., Australia, 2002.

[146] Hans-Jurgen Zepernick, and Adolf Finger, *Pseudo Random Signal Processing Theory and Application*, West Sussex, England: John Wiley and Sons, 2005.

[147] J. Deller Jr., J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals*, New York, USA: IEEE Press, 2000.

[148] Harry Levitt, "Noise reduction in hearing aids: A review," *Journal of Rehabilitation Research and Development*, vol. 38, no. 1, pp. 111-121, Jan./Feb. 2001.

[149] Hadhami Issaoui, Aïcha Bouzid, and Noureddine Ellouze, "Comparison between Soft and Hard Thresholding on Selected Intrinsic Mode Selection," in *Proc. IEEE conference on Sciences of Electronics, Technologies of Information and telecommunications*, Sousse, Mar. 2012, pp. 712-715.

[150] Byung-Jun Yoon, and P. P. Vaidyanathan, "Wavelet-based denoising by customized thresholding," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, May 2004, pp. 925-928.

[151] Martin R., "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on speech and audio processing*, vol. 9, no. 5, pp. 504- 512, Jul. 2001.

[152] Boll S. F., "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 27, no. 2, pp. 113-120, Apr. 1979.

[153] Gong Y., "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261-291, Apr. 1995.

[154] Weaver J. B., Yansun X., Healy D. M. Jr., and Cromwell L. D., "Filtering noise from images with wavelet transforms," *Magnetic Resonance in Medicine*, vol. 21, no. 2, pp. 288-295, Oct. 1991.

[155] Saeed Ayat, Mohammad T. Manzuri, and Roohollah Dianat, "Wavelet Based Speech Enhancement Using a new Thresholding Algorithm," in *Proc. of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, Oct. 2004, pp. 238-241.

[156] Carl Taswel, "The What, How, and Why of Wavelet Shrinkage Denoising," Computational Toolsmiths, Stanford, Rep. CT-1998-09, pp.1- 11.

[157] Lawrence R. Rabiner, Marvin R. Sambur, and Carolyn E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23. no. 6, pp. 552-557, Dec. 1975.

[158] Maciej Niedzwiecki, and William A. Sethares, "Smoothing of discontinuous signals: The competitive approach," *IEEE Transactions on Signal Proessing*, vol. 43, no. 1, pp. 1-13, Jan. 1995.

[159] Sheau-Fang Lei, and Ying-Kai Tung, "Speech Enhancement for Non-stationary Noises by Wavelet Packet Transform and Adaptive Noise Estimation," in *Proc. of International Symposium on Intelligent Signal Processing and Communication Systems*, , Dec. 2005, pp. 41-44.

[160] Hesham Tolba, "A Time-Space Adapted Wavelet De-Noising Algorithm for Robust Automatic Speech Recognition in Low-SNR Environments," in *Proc. IEEE 46th Midwest Symposium on Circuits and Systems*, vol. 1, Cairo, Dec. 2003, pp. 311-314.

[161] A. Lallouani, M. Gabrea, and C.S. Gargour, "Wavelet Based Speech Enhancement Using Two Different Threshold-Based Denoising Algorithms," in *Proc. Canadian Conference on Electrical and Computer Engineering*, vol. 1, Canada, May 2004, pp. 315-318.

[162] Slavy G. Mihov, Ratcho M. Ivanov, and Angel N. Popov, "Denoising Speech Signals by Wavelet Transform," *Annual Journal of Electronics*, pp. 69-72, Jun. 2009.

[163] Okawa S., Boochieri E., and Potamianos A., "Multi-band speech recognition in noisy environment," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, Seattle, May 1998, pp. 641-644.

[164] Tibrewala S., and Hermansky H., "Sub-band based recognition of noisy speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, Munich, Apr. 1997, pp.1255-1258.

[165] Lior Rokach (2010). *Pattern Classification Using Ensemble Methods (Series in Machine Perception and Artificial Intelligence*, vol. 75, World Scientific Publishing Company. [Online]. Available: http://www.ebook.downappz.com/?page=download&id=bmo%3Fi%22&file=Pattern_Classification_Using_Ensemble_Methods_Series_in_Machine_Perception_and_Artificial_Intelligence_.pdf.

[166] Isabelle Guyon, and Andre Elisseeff, "An Introduction to Feature Extraction," *Studies in Fuzziness and Soft Computing, Springer-Verlag*, vol. 207, pp. 1-25, 2006.

[167] Joseph P. Campbell, "Speech recognition: A Tutorial," *Proc. of the IEEE*, vol. 85, no. 9, pp. 1437-1462, Sept. 1997.

[168] C. J. Wild, and G. A. F. Seber, *Chance Encounters: A First Course in Data Analysis and Inference*, 1st ed. USA: Wiley, 1999.

[169] Peter Baxte, "Statistical Process Control Analysis Unraveled," *Distributive Management*, pp. 1-20, 2010. [Online]. Available: http://www.distributive.com/cms-assets/documents/9454-417659.spc-analysis-unraveled.pdf

[170] 68–95–99.7 rule. http://en.wikipedia.org/wiki/68–95–99.7_rule

[171] Joarder A. H., and Firozzaman M., "Quartiles for discrete data," *Teaching Statistics*, vol. 23, issue 3, pp. 86-89, Oct. 2001.

[172] Eric Langford, "Quartiles in Elementary Statistics," *Journal of Statistics Education*, vol. 14, no. 3, Nov. 2006.

[173] David Franklin, "Calculating the Quartile (or why are my Quartile answers different?)," in *Proc. NESUG*, Baltimore, Nov. 2007.

[174] Quartile. http://en.wikipedia.org/wiki/Quartile

[175] Jerrold H. Zar, *Biostatistical Analysis*, 5th ed. New Jersey, USA: Pearson Education, 2010.

[176] Gouda I. Salama, M. B. Abdelhalim, and Magdy Abd-elghany Zeid, "Breast Cancer Diagnosis on three different datasets using multi-classifiers," *International Journal of Computer and Information Technology*, vol. 1, issue 1, pp. 36-43, Sep. 2012.

[177] Robi Polikar, "Ensemble based systems in decision making," Feature in IEEE Circuits and Systems Magazine, pp. 21-45, 3rd quarter to 2006.

[178] Thomas G. Dietterich, "Ensemble Methods in Machine Learning," in *Proc. International Workshop on Multiple Classifier Systems*, Oregon USA, 2000, pp. 1-15.

[179] Zhongwei Zhang, Jiuyong Li, Hong Hu, and Hong Zhou, "A Robust Ensemble Classification Method Analysis," *Advances in Computational Biology Advances in Experimental Medicine and Biology*, vol. 680, 2010, pp. 149-155.

[180] David Gelbart, "Ensemble Feature Selection for Multi-Stream Automatic Speech Recognition," Ph.D. dissertation, Dept. of Eng. Elec. Eng. and Comp. Sci., Univ. of California, Berkeley, 2008.

[181] Saso Dzeroski, and Bernard Zenko, "Is Combining Classifiers with Stacking Better than Selecting the Best One?" *Machine Learning*, vol. 54, pp. 255–273, 2004.

[182] Lior Rokach, "Ensemble Methods for Classifiers," in *Data Mining and Knowledge Discovery Handbook*, US: Springer US, 2010, ch. 45, pp. 957-980.

[183] Shiliang Sun, Changshui Zhang, and Dan Zhang, "An experimental evaluation of ensemble methods for EEG signal classification," *Pattern Recognition Letters*, vol. 28, issue 15, pp. 2157-2163, Nov. 2007.

[184] David Opitz, and Richard Maclin, "Popular Ensemble Methods: An Empirical Study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169-198, 1999.

[185] Donn Morrison, and Liyanage C. De Silva, "Voting ensembles for spoken affect classification," *Journal of Network and Computer Applications*, vol. 30, issue 4, pp. 1356-1365, Nov. 2007.

[186] Smita Vemulapalli, Xiaoqiang Luo, John F. Pitrelli, and Imed Zitouni, "Using Bagging and Boosting Techniques for Improving Coreference Resolution," *INFORMATICA*, vol. 34, pp. 111-118, 2010.

[187] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[188] E. Bauer, and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting and variants," *Machine Learning*, vol. 36, no. 1-2, pp.105-139, 1999.

[189] Prem Melville, Nishit Shah, Lilyana Mihalkova, and Raymond J. Mooney, "Experiments on Ensembles with Missing and Noisy Data," in *Proc. 5th International Workshop on Multiple Classifier Systems (MCS)*, LNCS vol. 3077, Cagliari, June 2004, pp. 293-302.

[190] Kristína Machová, František Barčák, Peter Bednár, "A Bagging Method using Decision Trees in the Role of Base Classifiers," *Acta Polytechnica Hungarica*, vol. 3, no. 2, pp. 121- 132, 2006.

[191] Pengyi Yang, Yee Hwa Yang, Bing B. Zhouand, and Albert Y. Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, issue 4, pp. 296-308, 2010.

[192] R. E. Schapire, Yoav Freund, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771-780, 1999.

[193] Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Burlington, Massachusetts: Morgan Kaufmann, 2011.

[194] Hui-Lan Luo, Zhong-Ping Liu, "A Review of Ensemble Method," *International Conference on Affective Computing and Intelligent Interaction Lecture Notes in Information Technology*, vol. 10, 2012, pp. 22-26.

[195] Gavin Brown, "Ensemble Learning," *Encyclopaedia of Machine Learning*, Chapter No 00400, Springer, pp. 1-9, Jan. 2010.

Papers in International Journals

[1] Sonia Sunny, David Peter S., and K. Poulose Jacob, *"*An Improved Hybrid Feature Extraction Technique for Speaker Independent Isolated Words in Malayalam," *AASRI Procedia, ELSEVIER,* 2013. (Accepted for Publication)

[2] Sonia Sunny, David Peter S., and K. Poulose Jacob, *"*A New Algorithm for Adaptive Smoothing of Signals in Speech Enhancement," *IERI Procedia, ELSEVIER,* 2013. (Accepted for Publication)

[3] Sonia Sunny, David Peter S., and K. Poulose Jacob, "Design of a Novel Hybrid Algorithm for Improved Speech Recognition with Support Vector Machines Classifier," *International Journal of Emerging Technology and Advanced Engineering,* vol. 3, Issue 6, pp. 249-254, June 2013.

[4] Sonia Sunny, David Peter S., and K. Poulose Jacob, "Performance of Different Classifiers in Speech Recognition," *International Journal of Research in Engineering and Technology,* vol. 2, no. 4, pp. 590 – 597, April 2013.

[5] Sonia Sunny, David Peter S., and K. Poulose Jacob, "A Comparative Study of Wavelet based Feature Extraction Techniques in Recognizing Isolated Spoken Words," *International Journal of Signal Processing Systems,* vol. 1, no. 1, pp. 49-53, June 2013.

[6] Sonia Sunny, David Peter S., and K. Poulose Jacob, "Performance Analysis of Different Wavelet Families in Recognizing Speech," *International Journal of Engineering Trends and Technology,* vol. 4, no. 4, pp. 512-517, April 2013.

[7] Sonia Sunny, David Peter S., and K. Poulose Jacob, "Development of a Speech Recognition System for Speaker Independent Isolated Malayalam Words," *International Journal of Computer Science and Engineering Technology*, vol. 3, no. 4, pp. 69-75, April 2012.

[8] Sonia Sunny, David Peter S., and K. Poulose Jacob, "Optimal Daubechies Wavelets for Recognizing Isolated Spoken Words with Artificial Neural Networks Classifier," *International Journal of Wisdom Based Computing,* vol. 2, no. 1, pp. 35-41, April 2012.

[9] Sonia Sunny, David Peter S., and K. Poulose Jacob, "Recognition of Speech Signals: An Experimental Comparison of Linear Predictive Coding and Discrete Wavelet Transforms," *International Journal of Engineering Science and Technology,* vol. 4, no. 4, pp. 1594-1601, April 2012.

[10] Sonia Sunny, David Peter S., and K. Poulose Jacob, "Discrete Wavelet Transforms and Artificial Neural Networks for Recognition of Isolated Spoken Words", *International Journal of Computer Applications,* vol. 38, no. 9, pp. 9-13, January 2012.

[11] Sonia Sunny, David Peter S., and K. Poulose Jacob, "Wavelet Packet Decomposition and Artificial Neural Networks based Recognition of Spoken Digits", *International Journal of Machine Intelligence, v*ol. 3, Issue 4, pp. 318-321, December 2011.

[12] Sonia Sunny, David Peter S., and K. Poulose Jacob, "A Wavelet Based Recognition System for Malayalam Vowels using Artificial Neural Networks," *International Journal of Computational Linguistics Research,* vol. 1, no. 2, pp. 81-87, June 2010.

Papers in International Conferences

[1] Sonia Sunny, David Peter S., and K. Poulose Jacob, "Adaptive Smoothing and Wavelet Denoising for an Enhanced Speech Recognition System," in *Proc. Seventeeth International Conference on Image Processing, Computer Vision and pattern Recognition (The 2013 World Congress in Computer Science , Computer Engineering and Applied Computing),* USA, July 2013, pp. 795-800.

[2] Sonia Sunny, David Peter S., and K. Poulose Jacob, "Combined Feature Extraction Techniques and Naive Bayes Classifier for Speech Recognition," in *Proc. Third International Conference on Advances in Computing & Information Technology, (ACITY 2012),* Chennai, July 2013, pp. 155-163.

[3] Sonia Sunny, David Peter S., and K. Poulose Jacob, "A Comparative Study of Parametric Coding and Wavelet Coding Based Feature Extraction Techniques in Recognizing Spoken Words," in *Proc. CUBE International Information Technology Conference and Exhibition,* Pune, September 2012, pp. 326- 331. (Available in ACM Digital Library)

[4] Sonia Sunny, David Peter S., and K. Poulose Jacob, "Feature Extraction Methods based on Linear Predictive Coding and Wavelet Packet Decomposition for Recognizing Spoken Words in

Malayalam," in *Proc. International Conference on Advances in Computing and Communications,* Cochin, August 2012, pp. 27-30. (Available in IEEE Xplore)

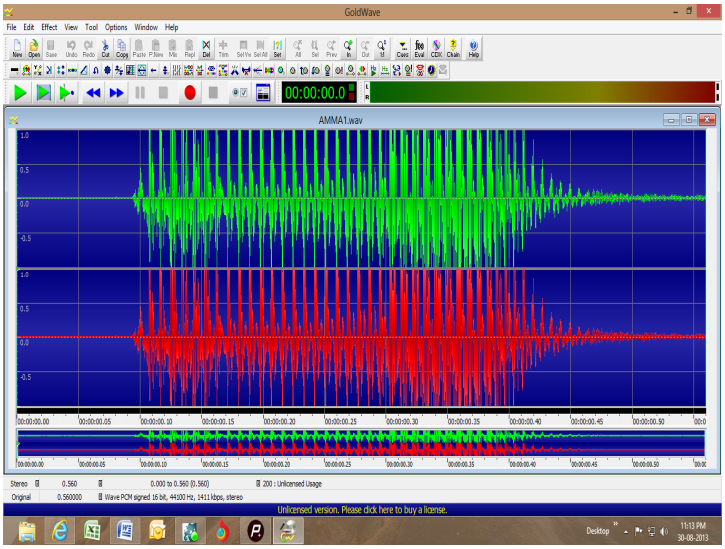
[5] Sonia Sunny, David Peter S., and K. Poulose Jacob, "Recognition of Spoken Digits: A Comparative Study of Discrete Wavelet Transforms and Wavelet Packet Decomposition with Artificial Neural Networks Classifier", in *Proc. of International Conference on Computer Science and IT Applications,* New Delhi, November 2011, pp. 71-75.

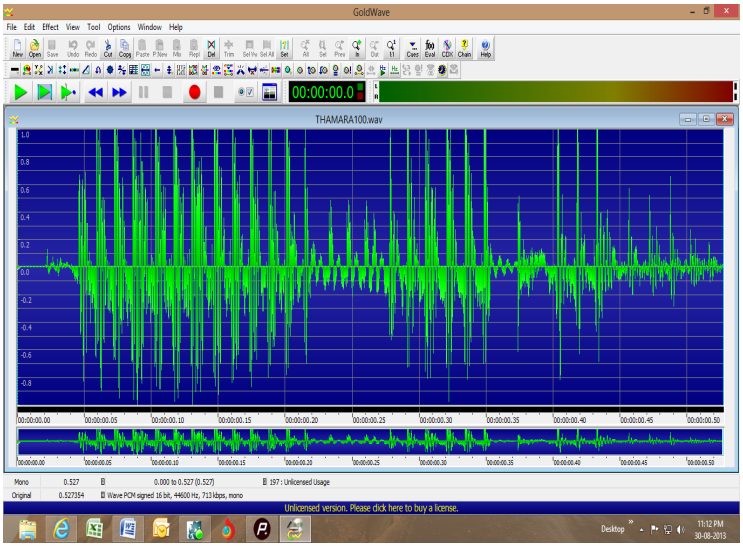
[6] Sonia Sunny, David Peter S., and K. Poulose Jacob, "Application of Discrete Wavelet Transforms and Artificial Neural Networks in Recognizing Spoken Digits," in *Proc. of the International Joint Colloquiums on Computer Electronics Electrical Mechanical and Civil,* Muvattupuzha, September 2011, pp. 71- 73.

APPENDIX

Appendix A

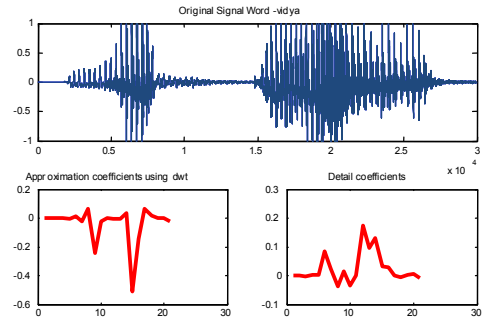
Sample Screen Shots from Goldwave



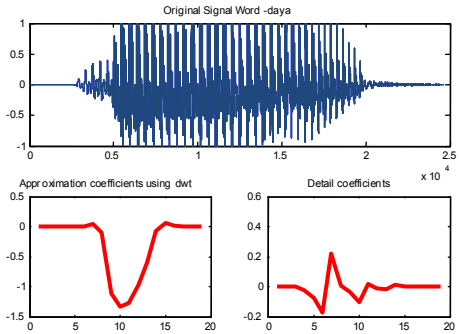


Appendix B

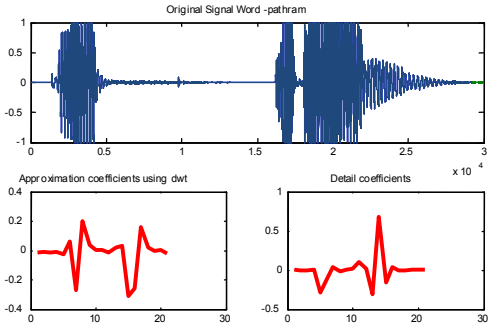
Sample Pictures of the 8th Level Decomposition of Isolated Words using DWT



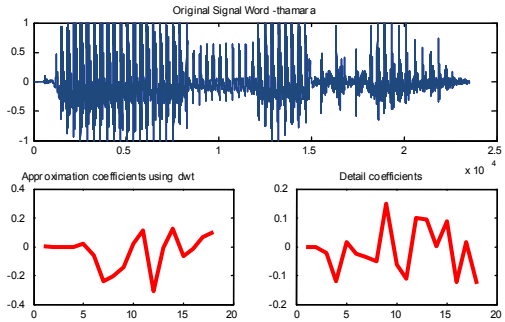
Decomposition of word 'vidya' വിദ്യ



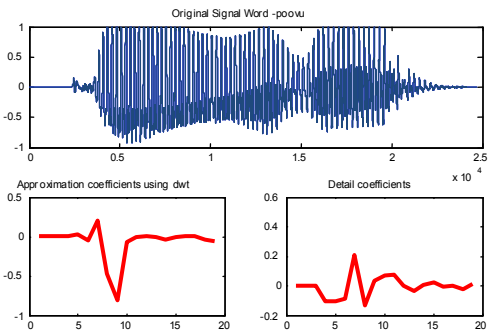
Decomposition of word 'daya' ദയ



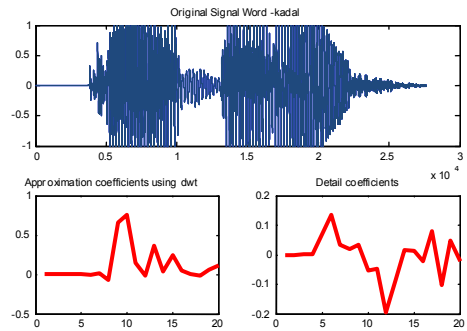
Decomposition of word 'pathram' പത്രം



Decomposition of word 'thamara' താമര



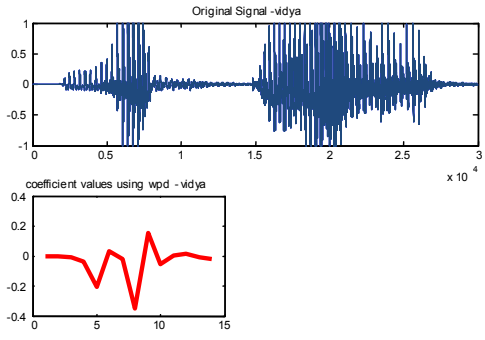
Decomposition of word 'poovu' പൂവ്



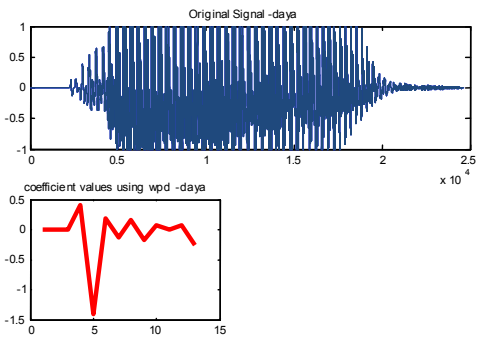
Decomposition of word 'kadal' കടൽ

Appendix C

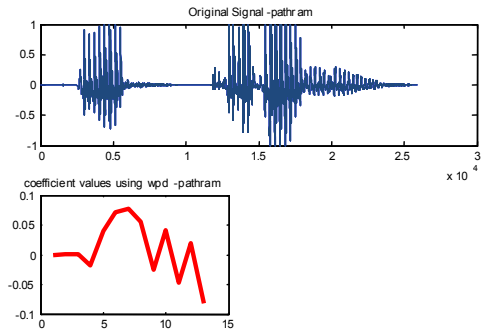
Sample Pictures of the 8th Level Decomposition of Isolated Words using WPD



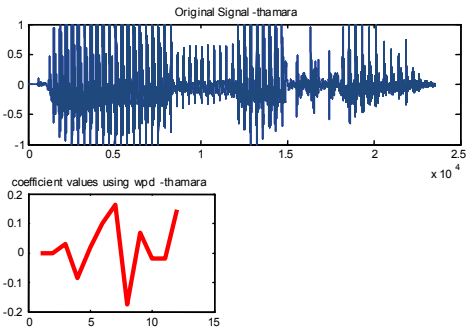
Decomposition of word 'vidya' വിദ്യ



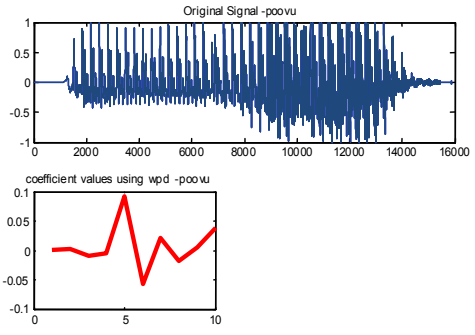
Decomposition of word 'daya' ദയ



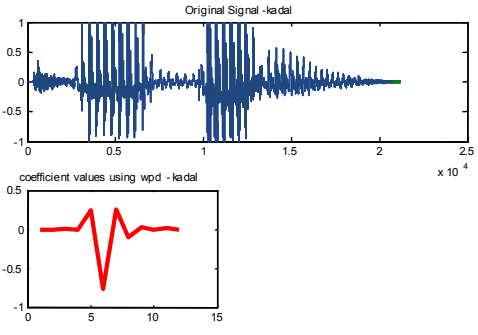
Decomposition of word 'pathram' പത്രം



Decomposition of word 'thamara' താമര



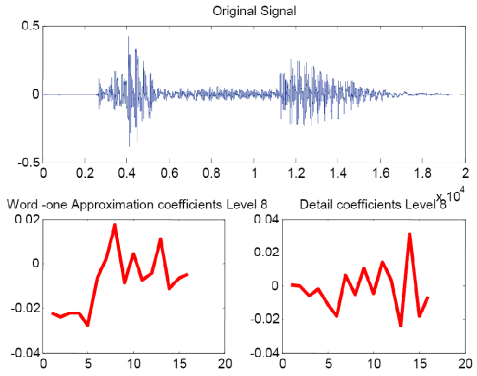
Decomposition of word 'poovu' പൂവ്



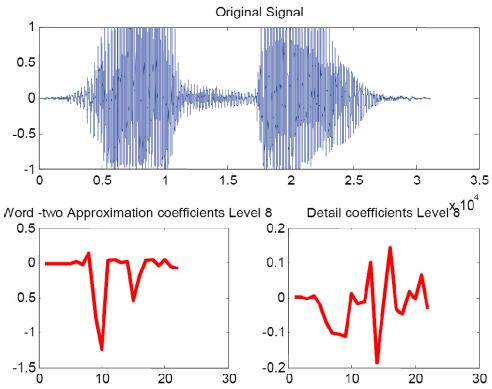
Decomposition of word 'kadal' കടൽ

Appendix D

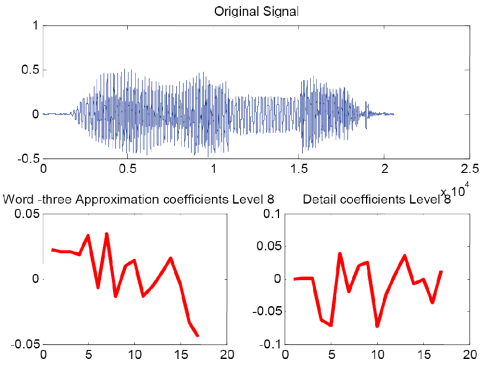
Sample Pictures of the 8th Level Decomposition of Digits using DWT



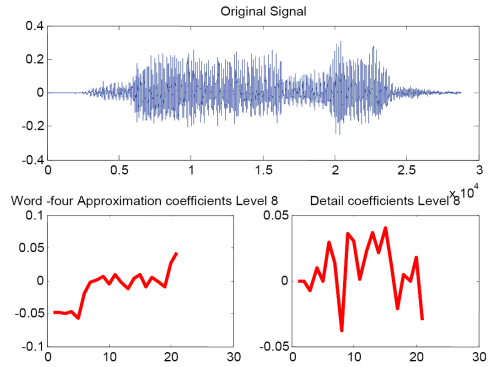
Decomposition of digit one ഒന്ന്



Decomposition of digit two രണ്ട്



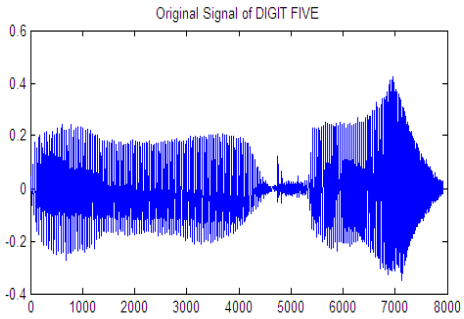
Decomposition of digit three മൂന്ന്

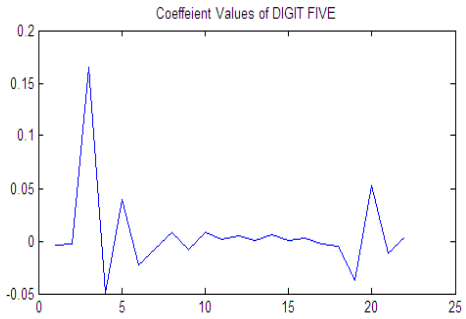


Decomposition of digit four നാല്

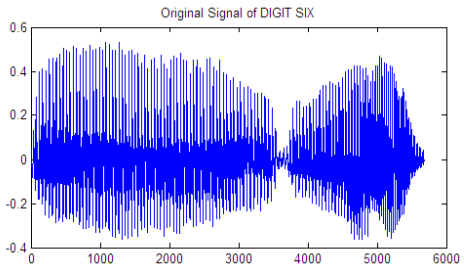
Appendix E

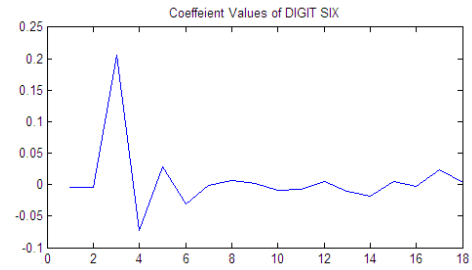
Sample Pictures of the 8th Level Decomposition of Digits using WPD



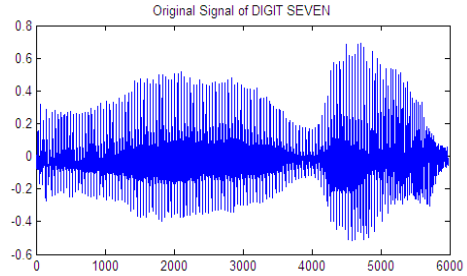


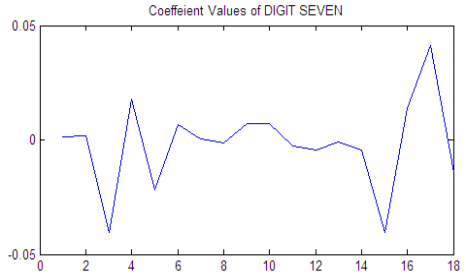
Decomposition of digit five അഞ്ച്



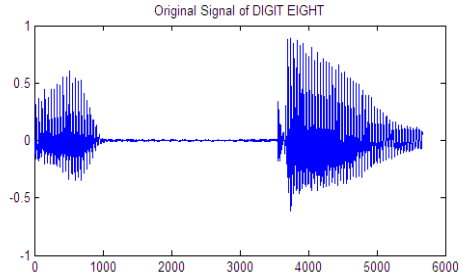


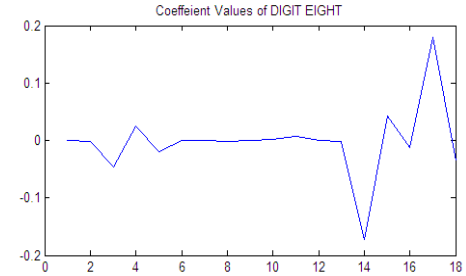
Decomposition of digit six ആറ്





Decomposition of digit seven ഏഴ്

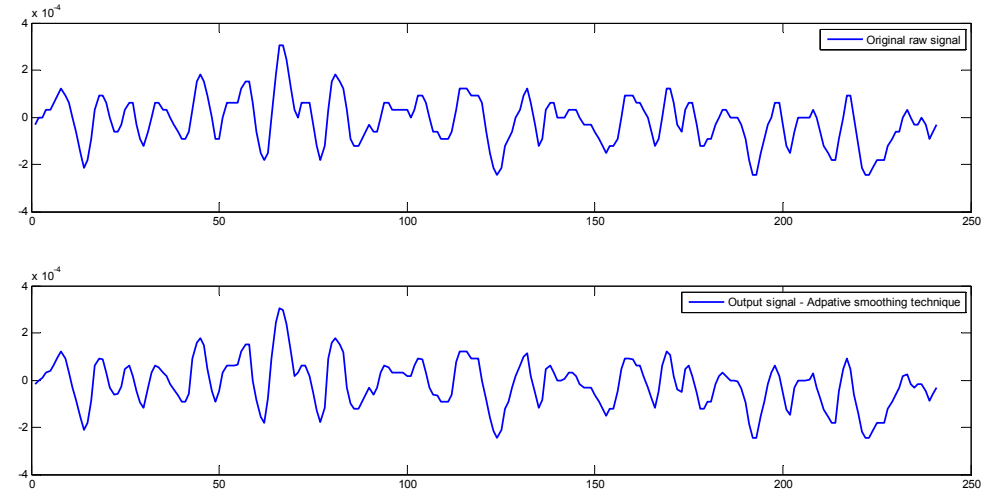




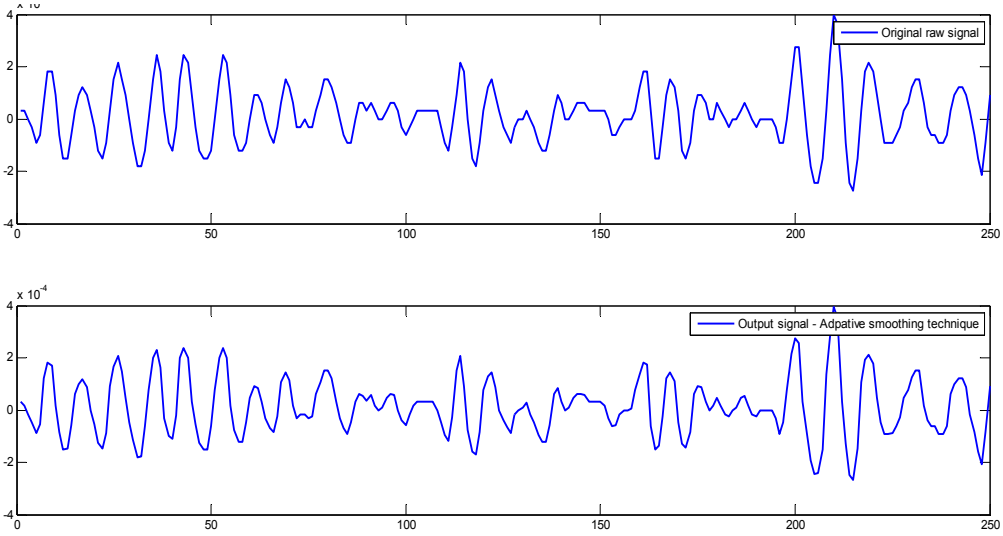
Decomposition of digit eight എട്ട്

Appendix F

Sample Plots of Signals before and after Adaptive Smoothing



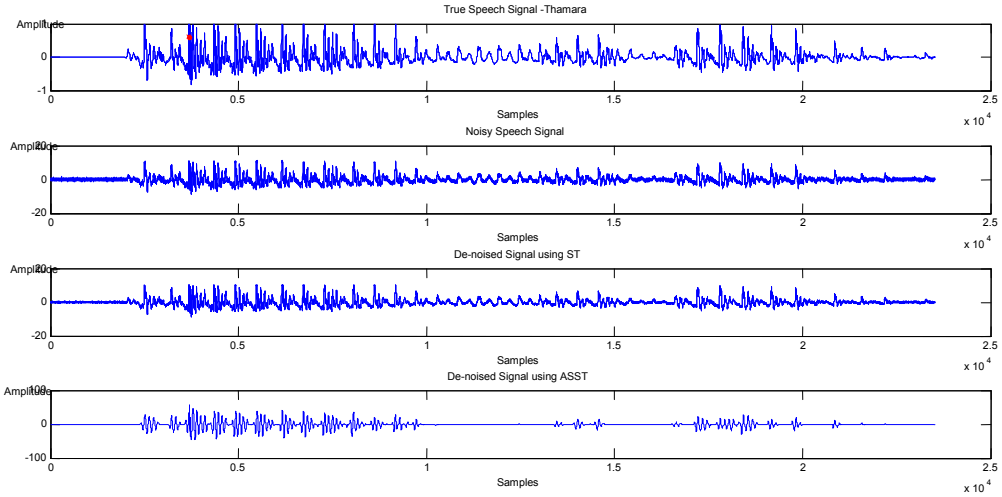
Part of the signal before and after adaptive smoothing



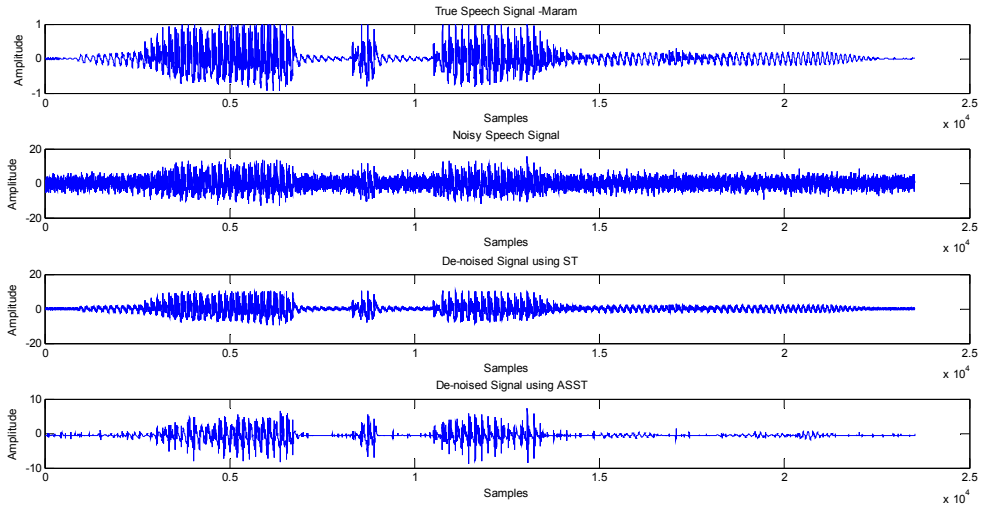
Part of the signal before and after adaptive smoothing

Appendix G

Sample Waveform Plots of Original Signal, Noised Signal With 5dB Noise, Denoised Signal using Soft Thresholding and Denoised Signal using Adaptive Smoothing Soft Thresholding



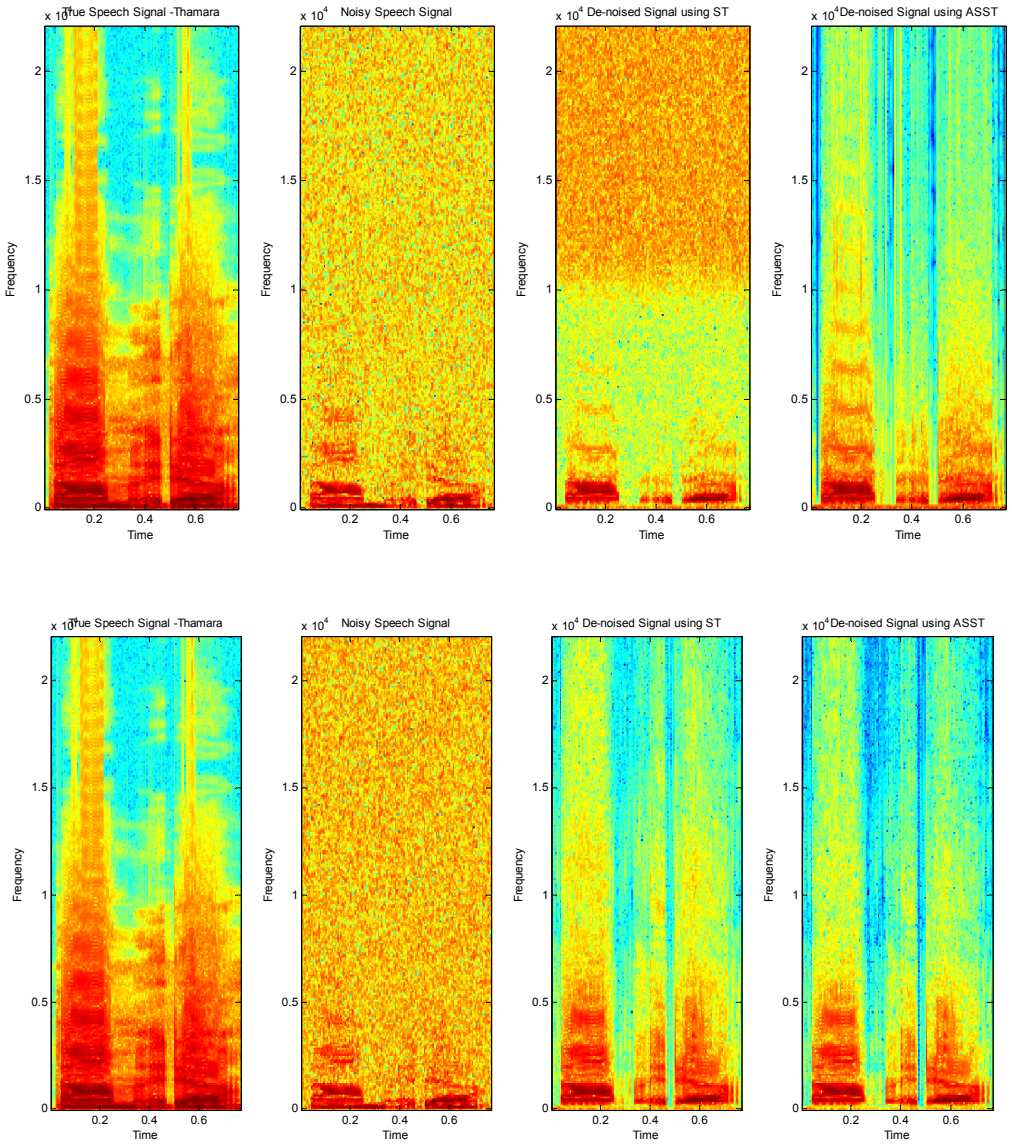
Waveform Plots of 'Thamara' താമര



Waveform Plots of 'Maram' മരം

Appendix H

Sample Spectrograms of Original Signal, Noised Signal with 5dB Noise, Denoised Signal using Soft Thresholding and Adaptive Smoothing Soft Thresholding



Spectrograms of word 'Thamara' തമര spoken by two different persons

Appendix I

Screenshots from WEKA