“SPEECH RECOGNITION & RECTIFICATION FOR ARTICULATORY HANDICAPPED PEOPLE”

A THESIS SUBMITTED TO

SAVITRIBAI PHULE PUNE UNIVERSITY
FOR AWARD OF DEGREE OF DOCTOR OF PHILOSOPHY
(Ph.D.)

IN THE FACULTY OF
ELECTRONICS AND TELECOMMUNICATION ENGINEERING

SUBMITTED BY
SMT. SANJIVANI SHANTARAM BHABAD
(PGS/2558)

UNDER THE GUIDANCE OF
PROF. (DR.) GAJANAN K. KHARATE

RESEARCH CENTRE

DEPARTMENT OF ELECTRONICS AND TELECOMMUNICATION ENGINEERING
MATOSHRI COLLEGE OF ENGINEERING AND RESEARCH CENTRE, EKLAHRE, NASHIK-422105, MAHARASHTRA (INDIA).
DEC 2019

CERTIFICATE OF THE GUIDE

Certified that the work incorporated in the thesis "**Speech Recognition & Rectification for Articulatory Handicapped People**", submitted by **Smt. Sanjivani Shantaram Bhabad** was carried out by the candidate under my supervision and guidance. Such material as has been obtained from other sources has been duly acknowledged in the thesis.

Dr. Gajanan Kashiram Kharate

Research Guide
Principal,
Matoshri College of Engineering and
Research Centre, Dist -Nashik

Date:

Place:

DECLARATION BY THE CANDIDATE

I declare that the thesis entitled **Speech Recognition & Rectification for Articulatory Handicapped People**", submitted by me for the degree of **Doctor of Philosophy** is the record of work carried out by me during the period from **12/07/2012 to 06/12/2019** under the guidance of **Dr. Gajanan Kashiram Kharate** and has not formed the basis for the award of any degree, diploma, associate ship, fellowship, titles in this or any other University or other institution of higher learning.

I further declare that the material obtained from other sources has been duly acknowledged and credited in the thesis.

Date:

Place:

Smt. Sanjivani S. Bhabad

Research Scholar
Matoshri College of Engineering and
Research Center, Nasik.

ACKNOWLEDGMENT

First and foremost, I would like to give all the credit to the Almighty for giving me the strength to overcome the difficulties with ease.

I am highly indebted to my guide ***Prof. Dr. G. K. Kharate***, for his motivation, exemplary guidance, valuable suggestions, and support through the whole course of my research.

I would like to thank from the bottom of my heart, ***Prof. Dr. J. J. Chopade*** Research Coordinator, E&TC Engineering***, Prof. D. D. Dighe***, Head, E&TC Department and all faculty members of the Matoshri College of Engineering, Nashik.

I sincerely appreciate valuable support and encouragement from ***Dr. V. J. Gond***, ***Dr. S. T. Gandhe and Dr. V.H. Patil.***

I offer my sincere appreciation for the management of K.K.Wagh College of Engineering education and research, Nashik, for providing me a learning opportunity.

I would like to thank Principal ***Prof. Dr. K. N. Nandurkar*** and HOD (E&TC) ***Prof. Dr. D. M. Chandwadkar***, of KKWIEER, Nashik, for their cooperation during the period of my research and my colleagues at KKWIEER, Nashik, who have willingly helped me in completing this work.

I am greatly indebted to my father Late ***Mr.shantaram S .Bhabad,*** my mother ***Smt. Sulochana S.Bhabad***, my mother in law ***Smt.Candabai D.Munot***, my brothers and sister.

This thesis is dedicated to my husband ***Pravin***, my sons ***Ashutosh*** and ***Ajinkya***, my daughters ***Kirti, Vishakha*** and ***Snehal***. Without their support and encouragement, there would never be any chance for this thesis to happen. The blessings, help and guidance given by my family and friends, time to time shall carry me a long way in the journey of life on which I am about to embark.

ABSTRACT

Approximately 1.4 per cent of human beings are unable to express their views because of the issue of articulation in order to improve the quality of life. Articulatory handicapped people are unable to use their ability for leisure in learning and personality development. The three types of speech impairments are articulation disorders, fluent disorders, and voice disorders. Due to various anatomical or physiological limitations in the production of skeletal, muscular or neuromuscular, errors in the production of speech sound occur. These disorders include: Omissions: (te for 10), Replacements: (cheven for 7), Distortions: (tree for 3).

A new form of speech recognition and rectification program is introduced in this study which identifies and rectifies the disordered speech of people with articulatory disabilities. The goal of this research is to develop algorithms for speech recognition and rectification to improve the communication skills of people with articulatory deficiencies. The database used for experimental assessment consists of a total of 1100 samples of digit zero to digit 10 that have an articulation problem. Since the types of articulation disorders are different for each individual so that no standard database is available, it is a challenge to collect this type of database. It is a time-consuming process to record such data as those suffering from articulation problems. In this research, the data is collected from 10 different users who suffer from these problems.

In the first part of the research, speech recognition is performed without phoneme separation by means of a feature extraction technique such as MFCC, LPC and RASTA PLP along with minimum distance Euclidean classifier. The experimental result shows that MFCC performs well among all these techniques. Speech recognition was achieved using various classifiers such as SVM, k-NN, ANN and HMM. The algorithm was developed using MATLAB 2017b and is being tested on a database generated for articulatory handicapped people. Various parameters related to the performance of the MFCC system have been measured. Appropriate parameters set relevant to MFCC are chosen so that the device must provide accurate speech recognition accuracy. The system performance was checked for different classifiers such as k-NN, ANN, SVM and HMM on the same set of MFCC parameters.

Compared to other algorithms that are implemented, MFCC with k-NN performs well. The k-NN classifier with Euclidean distance provides good results.

In the second part, the technique of phoneme segmentation is used to rectify disordered speech that improves accuracy of recognition. Speech on the basis of the lexicon table is divided into 2,3,4 and 5 segments. The MFCC features of each segment are extracted using the same parameter set. To predict the correct word, the classifier k-NN is used. Using the proposed algorithm known as the positive position search algorithm, correlation and occurrence of phoneme are found. Segmentation of the phoneme is more likely to be successful in speech rectification as it is easier to distinguish phonemes. The proposed algorithm has been found to be useful in improving speech quality and recognition scores.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBRIVATION

MFCC	Mel Frequency Cepstral Coefficient
HMM	Hidden Markov Model
ANN	Artificial Neural Network
SVM	Support Vector Machine
k-NN	K Nearest Neighbor
LPC	Linear Predictive Coding
RASTA PLP	Relative spectral Perceptual Linear Prediction
WCPL	Worst Case Processing Loss
ENB	Equivalent Noise Bandwidth
FFT	Fast Fourier Transform

LIST OF SYMBOLS

F Fundamental frequency

F1 First formant

F2 Second formant

F3 Third formant

F4 Forth formant

F_n n th formant

sign (x) signum function

E_r band level energy

C_r Spectral centroid

F_0 Fundamental Frequency

R_r Spectral roll-off

F_r Spectral flux

$S_n(e^{j\omega})$ Short-Time Fourier Transform

$W(e^{jw})$ Frequency response of window function

$\delta(\omega)$ Frequency response of impulse function

ω_c Cut-off frequency

F_s Sampling Frequency

t_w window length

Δf Frequency resolution

ΔL Frame shift

Hz Unit of frequency

f Speech input frequency

∇f Gradient vector

Hf Hessian matrix

λ damping factor

INTRODUCTION

1.1 Overview

To express feelings, thoughts, information related to real world and for daily routine work proper communication via speech signal is very important task. Speech is the most natural form of human verbal communication. Recognition of speech technology has enabled computers to follow commands of human voice and to understand human languages. The primary objective of speech recognition is to develop techniques and systems for the understanding of speech and natural language. Now days, it is used as an input to the machine for effective reorganization and processing.

The performance of automatic speech recognizers has improved a lot in past decade. In regards to the needs of the articulatory disabled population; most researchers have tried the existing technology. Research shows that although speech recognition applications for articulatory handicapped people are well within the capacity of available technology, it is primarily a lack of human factors work which impeding developments in this field. Several issues related to human factors are identified. The most promising areas for the application of speech recognition are in helping articulatory handicapped people [1].

The accuracy of automatic speech recognition for people with articulatory disabilities remains one of the major research challenges even after decades of research and development, The design of an automatic recognition of speech system requires careful attention to the issues viz. Speech class concept, speech representation, feature extraction methods, speech classification, database and performance assessment.

Spoken input is used to activate action and speech recognition when speaking to your computer and phone or app. All input forms are replaced with speech recognition, such as texting, clicking or choosing. It is a means to make devices and software more user-friendly and to increase productivity. A computer or program's ability to identify spoken language words and phrases and translate them to a machine-readable format called speech recognition. Accuracy and speed are calculated to calculate its

efficiency. Precision is calculated with the frequency of word error. WER operates at word level and detects transcription inaccuracies, although it cannot decide how the error occurred. Speed is assessed with a real-time factor. Speed recognition is affected by a variety of factors, such as pronunciation, accent, background noise, pitch and volume [2].

Speech disabilities affect the way a person forms speech. Stuttering is one of the most commonly experienced speech disorders. Many people with speech disabilities are unable to express their thoughts because of the issue of articulation even though they are mindful of what they want to say.

Voice recognition is studying voice signals and processing methods. The signals are usually interpreted in a digital format, and voice processing can be viewed as a specific case of digital signal processing for speech signals. The speech processing elements include the collection, control, transport, transmission and distribution of voice signals. The study of speech signals and their processing techniques was represented by speech processing and the study of the intersection of signal processing and processing of natural human language.

Digital voice identification, spoken word dialog systems, text-to-speak and automatic speech recognition are used for speech processing technologies. Data can also be taken out of speech (such as nationality, gender, language identity, or speech recognition). Speech may be a more convenient way of accessing, manipulating, and communicating information, but viable alternatives may exist: speech is not inherently the "normal" way to interact with a machine. Speech is hand-free, eye-free, fast and intuitive. It would be easier to talk about processing when there were simple linear relationships between acoustics and articulations, acoustics and perception. The transcription of speech plays an important role in automated speech synthesis and recognition. When speech is disordered, there will be no effective interaction between person to person or computer to person. When speech is disordered, the proper interaction between person to person and person to device will not be treated.

1.1.1 Form of Impairment of Speech

- **Phonological/articulation**

A standard development of speech sounds with substitutes, omissions, additions or distortions that may interfere with intelligibility [9].

- **Fluency**

A typical development of speech sounds with substitutes, omissions, additions or distortions that can impair intelligibility.

- **Voice**

It is characterized by abnormal vocal quality, pitch, loudness, resonance and/or duration of production and/or absence, depending on the age and/or sex of the individual.

- **Cleft lip palate**

An upper lip opening or breaking, the mouth roof (palate) or both

- **Spoken language disability**

Individuals with language problems have a profile because of their current level of language processing and functioning in language-related areas, including hearing, mental and speech skills.

- **Cerebral palsy**

It is the movement, muscle tone, or posture dysfunction that is caused by an insult to the immature in developing brain before birth.

- **Hearing disorders**

The physiological auditory system's impaired auditory sensitivity leads to hearing disorder. This disorder may limit speech and/or language development, understanding, production, and/or maintenance.

The thorough literature survey about speech recognition for speech disorder people has been carried out. The papers presented by the researchers in 1970, 1972 to 2018 papers related to the research work have been studied. The researchers have work on speech recognition but not on speech rectification for articulatory handicapped people. The detail literature survey is presented in Chapter 2.

1.2 Motivation

Communication is human life's nature. It is also the most powerful tool for dealing with everyday life. Verbal communication is special to individuals and is said to be the most powerful means of communication. The spectrum of communication issues includes speech, hearing and thought difficulties such as voice disorders, phonological disorders, fluidity disorders (stuttering or cluttering), language

disorders, impaired and/or arrested speech and language development due to hearing impairment, mental retardation and other problems.

Dictation is the most common use of speech recognition, where transcription of speech can be used to write letters / e-mails and other documents. This dictation technology has been widely used in many languages for several years and works very well. In many applications where speech recognition is hidden from the user, for example, in automatic transcription of audio archives, the recognition of coherent natural speech can be used in addition to dictation, in order to allow better organization and indexing. Speech recognition is also an important part of automated speech translation, which is still a very scientific technology. For automatic dialog systems, speech recognition is also used. These typically automatic telephone response systems contain information about a particular domain and can be interacted for natural speech. The speech of a user is detected, analyzed, appropriate queries are made or activities are performed, and a reply is produced, which is then transformed with the speech synthesizer into speech again. Examples of such dialog systems are Google and Apple's Siri.

The ultimate goal of researchers in this area is to establish an effective speech recognition and rectification system for people with articulatory disabilities. The various algorithms have been proposed and developed by many researchers. Research has been vigorously conducted over the past four decades in the area of speech recognition and tremendous progress has been made. Encouraging findings have been obtained and when working under different restricted conditions, existing speech recognition systems have achieved a specific degree of recognition accuracy. However, they are not working on the correction of the speech. In the case of articulatory disabled people, speech recognition is not sufficient, but correction is also required for proper communication in daily life. In this proposed system, we focus on speech recognition as well as speech rectification for articulatory handicapped person with disabilities through the use of phoneme separation techniques.

1.3 Problem Statement

The title of this research work is “Speech Recognition & Rectification for Articulatory Handicapped People”.

1.4 Objectives and Goals of Thesis

In the process of analysis and comparative study, it has been experimented whether a combination of some of these techniques can help in removing the drawbacks of individual methods of speech recognition for articulatory handicapped people. The prime objective is to devise a new approach of speech recognition and rectification for articulatory handicapped people.

The goal of the system is to enhance the abnormal person speech recognition technology, to make the technology available for the development of new applications and to improve the existing applications.

When improving the speech recognition technology the main focus is on the aspects which currently have relatively low quality.

The objectives of research work are:

- To understand speech recognition methods/techniques
- To understand speech problems of articulatory handicapped people
- To propose novel approach of speech recognition and rectification for articulatory handicapped people.
- To implement the proposed approach and carry out the experimentation for verification of results.
- To compare the performance of proposed technique.

  The primary objective of this proposed system is to use an appropriate feature extraction technique and an appropriate classifier to construct a complete ASR system. To achieve this goal, all speech recognition technology and algorithm processes need to be studied and reviewed in detail in order to gain a better understanding of the system.

1.5 Research Contributions

Following are the key contributions made by the research work while designing speech recognition and rectification for articulatory handicapped people.

- In case of speech recognition system the first step is speech acquisition, which mainly consists of three significant components such as sample rate, bits per sample and number of channels. The literature review in this work reveals that the researchers have used mainly sample rate of 16 KHz .This work gives better accuracy for sampling rate of 44.1 KHz with 16 bits per sample because for this sampling rate high frequencies were not sufficiently attenuated and aliased back into the audible frequency range.
- High frequencies give the acoustic model with more information; pre-emphasis filter increases energy in high frequencies and improves recognition performance. Many researchers uses first order Finite Impulse Response (FIR) high pass filter as a pre-emphasis filter. This work investigates the performance of pre-emphasis first order as well as second order filter with different filter coefficient. The finding in this work shows that, the filter order is not affecting on the performance of system. But performance of system is affected by coefficient of first order FIR filter.
- In this work different feature extraction techniques are used to investigate the performance of system namely MFCC, RASTA-PLP SPECTRAL, RASTA – PLP CEPSTRAL and LPC. The result shows that MFCC is most suitable technique for this work. By varying the different parameters of MFCC feature extraction technique, the performance of the system is tested and the same parameters were used for further work.
- Minimum Euclidean distance classifier, K-NN, SVM, ANN and HMM classifiers are used in this work to compare the system performance. The finding in this work shows that the recognition accuracy using K-NN classifier is better than other. The performance of K-NN classifier is depending upon the value of k as well as distance measurement function.
- This work is basically divided into two sections, first the speech recognition using without phoneme separation and second is using with phoneme separation. Phoneme separation is done in time domain using segmentation

technique. Features of segments are extracted by using MFCC technique and the performance of system is tested by using different classifiers such as K-NN, ANN and HMM. The result indicates that the recognition accuracy is more using K-NN classifier. Furthermore, this work investigates the effect of phoneme separation on the performance of the system. The proposed strategy shows improvement in prediction of word and the system performance. Rectification of word is done on the basis of finding correlation and occurrences of phonemes. The proposed algorithm (prediction using phoneme separation) exhibits better results than prediction of word without phoneme separation.

1.6 Scope of Project

Although different techniques have been proposed to recognize speech for articulatory disabled people, each of these techniques has few limitations, yet none of the techniques have been fully developed.

Existing systems and suggested methods for speech recognition of articulatory handicapped persons have scope for improvement. The most desirable need in existing systems is an efficient way to represent, store and retrieve the knowledge needed for natural conversation with minimum errors. While considerable progress has been made over the last few decades, there is sufficient scope for improving the system especially for people with articulatory disabilities and using the technology to benefit society. It is one of the most prominent areas for the research.

1.7 Organization of Thesis

The objective of this research work is to implement speech recognition and rectification system for Articulatory Handicapped People. The thesis is organized in five chapters.

Chapter1 gives an introduction and briefly describes different types of speech disorders. The types of disorders are explained in the context of the research. It also describes motivation, scope, objectives, and goals and gives an outline of thesis.

Chapter 2 discusses detailed study and literature review on existing feature extraction techniques, classifiers and phoneme separation methodology used. Findings of literature review are also highlighted.

Chapter 3 describes the physiology of speech production mechanism and processing of speech recognition. It describes the Fundamentals of speech production Mechanism, speech signal representation, human attributes, perceptual features, Different short time analysis used in speech recognition, Window Characteristics, Causes of Speech Disorder (Articulatory Handicapped), Types of Speech Impairments, Statistical Representation of Speech Recognition System and summary

Chapter 4 describes the development and implementation of speech recognition and rectification for articulatory handicapped people it provides mathematical model for speech recognition system Different feature extraction and classifier algorithms are developed and implemented using MATLAB 2017 B. It also describe proposed algorithm such as positive position such algorithm.

Chapter 5 provides detailed results of each algorithm produced along with a summary.

Chapter 6 Explain the conclusions and suggestions for further work drawn from the results.

Appendix A describes creation of database used in this research.

Prediction of correct word for articulatory handicapped people is a major component in the designing of the system. Therefore, the challenges in selecting suitable feature extraction techniques, classifiers need to be investigated and various techniques to diminish these challenges need to be analyzed.

The next chapter discusses about different techniques used for the same purpose.

Chapter 2

LITERATURE SURVEY

Speech impairment is a major barrier for people who suffer from it while making public presentations. There are different speech disabilities seen in adults and children.

Disorder of speech is an indicator of the problem with speech output, poor voice quality or pitch or volume problems or speech lag. In some peoples, several problems can be combined. Listeners may have trouble understanding what speech disorder people are trying to say. Most of the researchers have worked in this field.

The speech is a one-dimensional signal. In the digital age, voice signal is used as an input source for the machine. Considering the applications of voice signals in different areas, there are major challenges in developing an efficient and accurate system that supports speech recognition for articulatory disabled people.

2.1 Types of speech disorders

The National Center for Health Statistics Survey shows that 2.68 per cent of crore people suffer from various disabilities. Among these 7%, people have speech-related disabilities. Of these 7%, 56 percent of males and 44 percent of females have speech impairment. In 2015, Balaji V, G. Sadashiv appa [4], is reviewed different types of speech dysfunctions. Few of them are briefly explained as follows.

- **Cluttering**

In this disability the speaker delivers speech abnormally fast, irregular or both. This becomes unintelligible to the listener. For example, “Oh, I think I my speech is garbled. I speak too fast… but listener always says „What did you say to me? Speak slower please‟…. I hear I hear myself. Oh, my words get garbled to myself I garble my words…” Many experienced clinicians have managed only to cure a few persons with this disorder.

- **Dysarthria**

Due to complication in the muscles, there is dysarthria that helps people talk. This may be caused by the paralysis of larynx, lips, tongue, palate, and jaw in the physiological function. It is characterized by words that are poorly pronounced.

Other symptoms are, Creating tongue-ties sounds like mumbling, Whispering or Speaking very softly so hearing is very difficult, and Nasal voice / Breathy voice or stuffy, hoarse, strained, It become very difficult for corrections.

- **Lisp**

It is a difficulty in creating one or more specific speech sounds, for example, /s/, /z/, /r/, /l/ and 'th'. This is known as a Functional Speech Disorder (FSD). This type of speech obstruction with particular syllables was also called stigmatism. The cause of this problem is unknown, yet, it does not reduce the speaker's intelligibility very much.

- **Esophageal voice**

The patient may swallow or inject some air into his/her esophagus. The air in the esophagus then vibrates a muscle and creates esophageal voice. It is often difficult to learn and understand esophageal voice and people with this problem can only talk in short sentences with a quiet voice.

- **Stuttering**

This is also called as Stammering, a speech disorder in which sounds, syllables, or words are repeated or uttered for an extended duration than normal. People who stutter know what they're going to say, but have trouble saying it. The speaker experiences a sudden break in the flow of speech, and because of these behaviors, stuttering is feared to be obstructing in educational and working environment affecting performance, as well as stalling social communication.

Speech Sound Disorders (SSD) is typically a class of speech issues that includes voice–sound production disabilities. SSD includes several sub-categories ranging from mild articulation problems to serious phonological disorders involving multiple speech–sound production defects and decreased smartness. Other names used for SSD are "disability of speech" or "disability of speech." Subcategories include articulation disorder, phonological disorder, speech apraxia in infancy, and dysarthria. Disorders of speech sound can develop from childhood or after an injury to the stroke or chest. Cerebral paralysis is a brain injury that inhibits the function of the muscle and makes it sluggish. With cerebral palsy, the result of muscle tightness, repetitive movement, and speech loss is clearly visible. These individuals can use specifically trained speech recognition systems for their dialect and accents.

In 2009, Caroline Bowen [5] describes the definition, causes and difference of articulation disorder and phonological disorder. Articulation is the process according to him by which words and sounds are created when the lips, tongue, jaw, teeth and palate change the air from the vocal folds. The listener does not understand a person with an articulation disorder's words because they are unable to correctly construct the words or sound. This problem may be caused by physical conditions such as palate cleavage, a syndrome that causes difficulty in producing sounds / words, or hearing loss, or may be due to other problems in the mouth and cerebral palsy. Examples of articulation problems include replacing one sound with another (e.g., saying *ken* for *ten*), or leaving out sounds (e.g., *tree* instead of *three*) or adding sounds to words ("pinanio" for "piano") are examples of articulation errors. A peculiar change involving the letters "s" and "z" relates to the problem of listing. The person who lisps alters the sounds with "ch" ("seven" sounds like "cheven").

In phonological disorders (also known as *phonological process disorders*), the person develops set paradigms of sound errors. People with phonological disorders have hardship in learning the sound system of the language, and may not figure out that, changing sounds can change meanings. They generate consistent error patterns

Automatic speech recognition (ASR) based approach for speech therapeutics of aphasic patients was reviewed by Norezmi Jamal et.al. In 2017 [6].

Aphasia is loss of language functioning and having a problem to communicate orally resulting from a stroke or brain injury in the absence of sensory, motor, or cognitive wreckage. In this work it is estimated that from 21% to 38% of stroke patients endures from aphasia. Symptoms of aphasia are quite subjective, varying from one speaker to the other. Some people with aphasia know a word which might feel like having on tip of the tongue but it is just difficult for them to get the right words out. Inaccuracies in phonemic, misrepresentations of articulation, and speech dysfluencies in speech production is revealed in people with aphasia.

2.2 Techniques Used

Speech recognition for articulatory disorder is an ideal example of multidisciplinary research. In the present era, most of the research works focused on the automatic speech recognition for articulatory disorder and classification, by methods of acoustic analysis, feature extraction, neural network and statistical method. The following

section presents an overview of previous work in the literature, which focuses on how automatic speech recognition for articulatory disorder is practiced, how experiments are designed and their results assessed. Below is the generic diagram for Speech recognition system, below image helps to explain two important techniques' used i.e. feature extraction and classifier. Below section will explain how people have used different feature extraction technique and classifiers to get result for speech disorders.

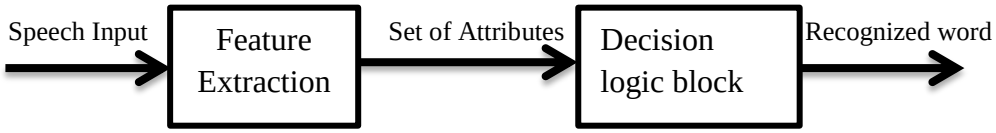


Figure 2.1: General Speech Recognition System

2.2.1 Artificial Neural Network (ANN)

In 1997, Selim S. Awad [7] presented a compelling and cost-effective speech therapy tool for those with speech stuttering disorder. Speech Fluency Treatment Tool is used; it takes the form of software combined with the computing power available on personal computers. The software works by clarifying to the stuttering client if his / her voice is erratic and conflicts with what his / her clinician recommends. Audio and visual signs are used to demonstrate where the client's voice is different from what the clinician expects. As a visual feedback to the user, the real-time showing of the average magnitude profile is equivalent to the spoken utterance of the client. The client must visually compare their average magnitude profile with that of the reference utterance and adjust their speech as necessary to suit the reference. Comparison is made between the origin, start and end positions, amplitude and length of the two average profiles of magnitude. A ranking is given in each of these classes, relating to the output of the company. To promote and retain speech fluency, customers learn to change their speech sound and word frequency, loudness, length and onset.

In 1995, P. Howell et al [8] used autocorrelation function and envelope parameters as the input vector of ANNs, providing a remarkable accuracy of around 80%. The same researchers [9, 10] used the fragmentation factor, spectral measures, component work length and energy to distinguish between fluent and dysfluent terms. The ANNs correctly identified 78.01 percent of the dysfluent (combination of extensions and repetitions) words.

It's Geetha et. Al.[11] conducted research in 2000 on the classification of childhood dysfluencies using ANNs. Difference between normal non-fluency and stuttering is obtained by the author on the basis of variables such as age, sex, type of dysfluence, frequency of dysfluence, duration, speech rate, historical, attitudinal and behavioral scores, family history and 92 percent accuracy.

In 2017, it was Jun Ren, et. Al.[12], the author embraced Deep Belief Neural Networks (DBNs) to model the distribution of dysarthric speech signals. The authors used the DBN model rather than the GMMs. The DBN system offers a computationally in-depth approach to speech recognition. The features of the vocal tract length normalization (VTLN) process extraction feature are used. The enhanced MFCC features are used to train the GMM-HMM model. The derivative and acceleration are also included in order to overcome the conditional independence hypothesis of the HMMs and the speaker adapted features are used to train the DNN. The work concludes that the use of DBNs makes the detector more resilient against data variations of speech signals generated by different degrees of severity of dysarthrical expression. The sentence error rate of this system is still a bit high.

S Reza Shahamirietal [13] in 2014, identifies the best-performing set of MFCC parameters that can serve as dysarthric acoustic features to be used in the ASR-based Artificial Neural Network (ANN). These researchers also study the use of ANNs as a fixed-length isolated-word SI ASR for individuals suffering from dysarthria. Best accuracy is achieved when speech detectors are trained by traditional 12 coefficient MFCC features without the use of delta and acceleration features shown. Word recognition rate of 68.38 percent is acknowledged the dysarthric speech in the proposed speaker independent ASR.

In 1995, Jayaram and Abdelhamied [14] researched the application of ANNs in a speaker-dependent dysarthric speech recognition system. Researchers provide two recognizers, the first having been trained using MFCC parameters and the second using the formant frequencies. The first system performed better than the second system. The ASR system has been trained and evaluated with Dysarthria using only one subject, so the results are not significant.

Word recognition based on stop-gaps detection, discerning vowel prolongations, syllable repetition detection is done by Czyzewski et al. in 2003[15]. The system

includes 6 specimens of fluent speech and 6 specimens of speech with stop-gaps. ANN classifier and rough set were proposed by the researchers to detect stuttering events. The results of the experiment show that better scores are obtained using a rough set-based system that gives 90 percent more accuracy than the ANNs with an average accuracy of 73.25 percent.

Szczurowska et. in 2006 [16] used neural networks to identify and categorize non-fluentand fluent samples of speech. The researchers used eight stuttering speakers ' spectral features. Such features were extended to the identification and classification of fluent and dysfluent by Kohonen and Multilayer Perceptron Networks. The authors achieved a best result of 76.67 per cent with the best network, assembled using 171 input neurons, 53 hidden layer neurons and 1 output neurons. A system related to an automatic detection of dysfluency in stuttered speech proposed in 2009, by Świetlicka, et al. [17]. The dataset used in this process consisted of different samples collected from 8 stuttering people, 59 fluent samples of speech and 59 non-fluent samples of speech. Speech samples were analyzed using twenty-one digital 1/3 octave center frequency filters between 100 Hz and 10000 Hz. These features of speech samples were used as inputs for the networks. The Multilayer Perceptron (MLP) and Radial Basis Function (RBF) networks are used to classify and recognize fluent and non-fluent speech samples. For each network, the classification accuracy is obtained from 88.1% to 94.9%. Research conducted has shown that artificial neural networks, particularly in the non-fluent, can be a useful tool in speech analysis. The first application of the neural network allowed the dimensions of the input signals to be reduced and made it possible to represent non-fluent speech by reflecting the syllabic structure of utterances and exposing dysfluent fragments. With regard to the criteria considered, the networks achieved small errors in all data groups and high accuracy, sensitivity and specificity values. Neural networks are a tool that could support research in the field of intelligent speech recognition systems. Networks, due to generalization, modeling and complicating structures, could also help to understand the principles accompanying non-fluent signals.

Based on the acoustic nature of the abnormal speech, Nayak. J, et al[18] proposed in 2005 a system that provides relevant information on the type of disorder in the speech production system. The signals detected are non-stationary and may contain

symptoms of current illness or warnings about impending illnesses. The indicators may be present at all times or may occur at random at certain intervals of the day. Studying and diagnosing abnormalities in volume data collected over several hours is exhausting and time consuming. Computer-based analytical tools can therefore be very useful in diagnosing intensive analysis and information identification over day-long intervals. The author used the artificial neural network to identify those diseases. Continuous wavelet transformation patterns. Discrete wavelet transform (DWT) coefficients have been adopted as a feature vector for classification. Using neural network technique, speech specimens are classified into three classes of normal viz, paralysis and hyper-function Continuous wavelet coefficients are used for the analysis of speech samples. In this research, the neural network classifier is presented as a diagnostic tool to assist the physician in the analysis of speech diseases. The accuracy of the classifier depends on a number of factors, such as the size and quality of the training set, the rigor of the training provided, and also the parameters chosen to represent the input. The classifiers presented are effective, giving approximately 80–85 percent accuracy. Analysis of the same signal through continuous time wavelet transformation of the speech signal can also provide a visual pattern that can be of significant diagnostic help. A finite length window is used at this point to locate a particular event along the time scale. This window, carrying along the signal in time, can successively estimate the spectral components of the signal. For the signal having, both slowly varying components and rapidly changing transient events, short time Fourier transform (STFT) fails. An infinite length window gives the FT, which gives perfect frequency resolution, but no information about time. The narrower length of the window gives better and better resolution of the time. Stationary windowing assumption provides lower frequency resolution. Transform Wavelet overcomes this issue. It uses small high-frequency windows and longer low-frequency windows. Discrete Wavelet Transform provides sufficient data for analysis as well as synthesis of the original signal, with a significant reduction in computation time.

In 2018, Paria Jamshid Lou et al.[19] conducted research on a simple, effective model for automatic detection of dysfluency, called an auto-correlative neural network (ACNN). The model uses a convolutionary neural network (CNN) and increases it at the deepest layer with a new auto-correlation operator capable of capturing the types

of "soft copy" dependencies typical of speech repair dysfluences. In experiments, the ACNN model performs better on a dysfluency detection task with a 5% increase in f-score compared to the CNN baseline, which is close to the prior best result on this task. Using the ACNN, this study achieves competitive results for the detection of dysfluency without relying on any hand-crafted features or other representations derived from pre-existing system output.

2.2.2 Hidden Markov Model

M. Wiśniewski, in 2007 et al.[20] recommended a system for implementing HMM classifier to recognize speech disorders prolonged fricative phonemes. The most common MFCC features are used as input to the classifier of HMM in this work. This feature provides too many parameters, but these cannot be decreased, as there may be a lack of important information, which may result in poor recognition effectiveness. The K-means algorithm is used to pick from the codebook the correct feature set. When solving two issues, the best identification effectiveness and the appropriate time required for calculations can be achieved: the selection of the HMM and the proper preparation of input data. Reducing the size of the codebook limits the proportion of acknowledgement. The influence on the acceptance of the number of states is of marginal significance. The experiment shows approximately 80% precision of identification. In 2000, in order to evaluate the degree of stuttering during therapy session, Nöth et al.[21] combined the work of Speech language Processing and speech recognition system. The program will carry out statistical analysis such as counting and classifying common repetitions, pauses and length of phonemes. The observable stuttering measures used to define the degree of stuttering are the occurrence of dysfluent portions of speech, dysfluent length, and speaking speed. The list consists of thirty-seven patients with signs of stuttering, either reading the whole passage or beginning the passage. The word and phoneme accuracies of the stuttered text display a correlation coefficient of up to 0.99 in relation to the number of observed dysfluences. In the future, the researcher perceived the results still raw and needed more experiments, especially with stutterers clearly either belonging to the type of repetition or blocking.

Fangxin Chen et.al[22] in 1997, examines the residual vocal ability of people with significant motor disability followed by extreme dysarthria. An adaptive word

detection algorithm is developed in this speech recognition system to detect words in highly irregular dysarthric speech. The Mel frequency cepstrum coefficients (MFCC) are used for parametric representation of the speech signal and the left-to-right discrete hidden Markov model (DHMM) is used for pattern recognition. The experimental results confirmed the system's high performance, particularly when large amounts of data were used for training (92%). Advance analysis is done over time to adapt the DHMM algorithm to the variation of the intra-speaker. Speech information with complete phonemic stock and more topics are also required to build a robust system with strong generalization. The system is feasible in its current form for applications of small vocabulary communication and control aids based on speech. Frank Rudzicz, in 2011[23], proposed an innovative method for acoustic-to-articulatory inversion to predict the positions of the acoustic vocal tract using a nonlinear Hammerstein system. The TORGO dysarthric articulation database was used for experimentation measuring points in the vocal tract. For most vocal tract variables, this access uses adaptive kernel canonical correlation analysis and is found to be significantly more accurate than mixed density networks, at or above the confidence level of 95 percent. In addition to this study, a new approach for ASR has been suggested in which acoustic-based theories are re-evaluated according to the probabilities of their task-dynamics articulatory realizations. The method integrates high-level, long-term voice development aspects and is found to be significantly more accurate than hidden Markov models, complex Bayesian networks, and Kalman filters switching.

In 2011, Frank Rudzicz et al.[24] enlisted the use of techniques for acoustic and lexical adjustment to account for articulatory errors made by dysarthric speakers. Consistent articulatory deviations in dysarthric speech can be exploited through speaker and pronunciation lexicon adaptation resulted in an average absolute word error reduction of 22.87 percent (42.11 percent relative), which is significant for the relatively large vocabulary size used in these experiments. Adaptation of the pronunciation lexicon showed statistically significant changes on models adapted to speakers and based on speakers. While the results obtained are heartening, in dysarthric speech, phonetic articulatory errors are only part of the problem. Due to the presence of excessive involuntary coughing, intermittent articulatory breakdowns,

prosodic disturbances, stuttering, and unintentional breaks in dysarthric voice, the role becomes more complicated.

In 2014, Speech Assistive Technology for dysarthric speech was developed by Caballero-Morales et.al[25]. This study is based on pronunciation pattern modeling. The system presents an approach that combines multiple paradigms of pronunciation to improve the recognition of dysarthric speech. This combination is accomplished by weighing an Automatic Speech Recognition (ASR) system's responses when setting different language model rules and limitations. Genetic Algorithm (GA) is used to predict each response's weight. GA also optimizes the implementation technique structure (Metamodels) based on discrete Hidden Markov Models (HMMs). The GA uses dynamic uniform mutation or crossover to further change the candidate weight and structure sets to enhance the Metamodels outcome. Using a larger vocabulary, sustained auditory tools from the Nemours dysarthric speech repository were implemented to check the method. ASR tests with the proposed approach on these resources showed accuracy of recognition over those obtained with standard metamodels and a well-used technique of adaptation of speakers. These results were statistically significant.

Multimodal speech recognition device proposed by Elham S. Salama et al.[26] in 2014 in speech disorder. The proposed system improves disordered speech's robustness performance. It is used on the basis of speech and visual elements for people with dysarthria speech disorder. The Mel-Frequency Cepstral Coefficients (MFCC) is used as the characteristics of the acoustic speech signal. Discrete Cosine Transform (DCT) Coefficients for the visual component are extracted from the mouth region of the speaker. The characteristics of the facial and mouth regions are extracted using the Viola-Jones algorithm. Then these output characteristics are concatenated on one vector. The classifier Hidden Markov Model (HMM) is used as a decision logic circuit on the acoustic and visual part mix function vector. The independent UA-Speech English server is used for research. The results of this research show that the multimodal system using visual features is more efficient and improves the accuracy of speaker-dependent experiments up to 7.91 percent and speaker-dependent experiments up to 3 percent.

Marek Wiśniewski and others in 2007[26] HMM's were used to recognize two kinds of speech issues in an acoustic signal: extension of fricative phonemes and blockades with stop phonemes repetition. In this study, test results of recognition effectiveness are accessible through HMM models within different configurations of considered speech disorders. There were overview models applied for identification of a category of illness, as well as models related to individual phoneme disturbance. The models developed for specific phonemes were used not to identify individual phoneme disruptions but to recognize a class of dysfluency. Speech disorders are often sounding that are known phonemes of entirely different shape and it is almost impossible to recognize individual phoneme's non-fluence. The recognition ratio of speech disorders with HMM use depends primarily on a very accurate teaching pattern selection. In the case of repeated blockades, the problem is a pause that occurs in a study. Sometimes this contributes to misinterpretation, such as the perception of silence as a disturbance. Another important conclusion is that it is crucial to properly select the window length. The longer sections lead to fewer variations and also to lower the number of incorrect recognitions. On the other hand, a too wide window selection leads to less sensitivity. A selection of a proper probability threshold is very important in the process of recognition. It is a trade-off between the level of sensitivity and predictability. Author suggested implementing procedures to pick the threshold automatically. Le et .al. in 2016[28] employed University of Michigan Aphasia Program (UMAP)database consisting of a large continuous English-language vocabulary collected from individuals with Aphasia (6 females, 11 males, 58 ± 14 years of age). The built-in microphone of the tablet with a sampling rate of 44.1 kHz is used to record audio. Using DNN-HMM with MFCC and LDA features, the authors attempted to enhance the ASR of Aphasic speech. Unfortunately, due to the data shortage problem in aphasic speech recognition, DNN-HMM did not yield a promising result. The error rate of using DNN-HMM was 42.9 percent, which is greater than the GMM-HMM method (39.7 percent) because the acoustic model with the extracted features is not that robust enough. There is a possibility that during the front-end process the features were distorted by noise.

Cantonese Aphasia Bank's server used by Lee et. Al.in 2016[29] for their research work. This list consists of spontaneous oral narratives recorded in a file for 149

unimpaired Cantonese native speakers and 104 individuals with post-stroke aphasia dependent on the talking function. A head-worn condenser microphone and a digital recorder with 44.1 kHz sampling speed were used to record the sound. They found that with MFCC and LDA features, GMM-HMM and DNN-HMM produced 58.2% and 57.8% of the error rate respectively. It concluded that acoustic models are not the most critical issue that produces the low precision that only has the smallest differences in error rate. In the reference[29], the authors ' error rate is higher than in the reference[28]. Speaking style is the factor, and language models are the main challenges in achieving high precision in ASR.

2.2.3 Support Vector Machine

Mark Johnson-Hasegawa, et.al.[30] for speakers with very poor intelligibility demonstrated automatic isolated digit recognition system in 2006. A variety of spastic dysarthria-related symptoms result in poor intelligibility. For two subjects, HMM-based digital recognition was achieved, but failed for the subject with the most pronounced tendency to reduce or delete all consonants in a word. On the contrary, digit recognition experiments were successful for two subjects, but not for the subject with a sluggish, deliberate stutter, using a fixed word length using SVM classifiers. Research shows that the HMM's dynamic time warping features provide some robustness against large-scale word-length variations, while the regularized discriminative error metric used to train the SVM gives it some robustness against consonant reduction and deletion.

In 2017,Carlos M. Travieso.et. al. [31] proposed a novel method for an automatic detection of voice diseases. The process used by the fisher evaluator to transform Discrete HMM (DHMM) to hyperdimensional. The RBF-SVM, which is trained in the K-fold cross-validation strategy, is used for classification. Results of approximately 99 percent accuracy are obtained for three different voice disease datasets. Linear methods or combinations of linear and nonlinear methods or continuous speech signals have been used to detect voice pathologies in sustained phonations. Three data sets of different diseases were used in this system: cleft lip and palate which produces hyper nasal voice, Parkinson's disease that produces dysarthrical expression, and laryngeal pathologies that produce dysphonia. The results indicate that the proposed approach appears to be appropriate and robust to detect all

these pathologies. The proposed approach is contrasted with other classification systems commonly used in the state of the art. In general, a specific identification, i.e. an RBF-SVM and an HMM-based classifier, is carried out. Both approaches showed a drop in accuracies compared to those obtained with the transformation based on DHMM. The main drawback of the methodology introduced here is the acoustic conditions of speech recordings. The researchers conducted preliminary tests with poor accuracy with recordings recorded under uncontrolled conditions. Another array of experiments are based on recordings of speech captured under noise-controlled conditions. JianglinWang.et.al, proposed various methods for classifying and detecting the voice of patients with vocal fold disorder using the pattern recognition method in 2007[32]. The HMM (Hidden Markov Model), the SVM (Support Vector Machine) and the GMM (Gaussian Mixture Model) are capable of extracting those data characteristics to create a model that fits the speech signal when the three methods are applied to the signal. Work was conducted to investigate the use of such models to estimate a speech signal's performance. The parameters of MFCC (Mel Frequency Cepstral Coefficients) are extracted from the mixed voice data set for the HMM. Six characteristic parameters were previously selected for GMM and SVM (Jitter, Shimmer, NHR (Noise-to-Harmonic Ratio), SPI (Soft Phonation Index), APQ (Amplitude Perturbation Quotient) and RAP (Relative Average Perturbation). To classify the mixed pathological voice data set, classifiers based on HMM, GMM and SVM were used. Although the absolute difference in the frequency is small, it still indicates that the GMM can be used in pathological voice classification as a more reliable classifier, giving us a comparable classification rate to the ANNThe best cases of train data and test data in the GMM-based method are 4.6 percent and 3.5 percent better than the HMM-based method. However, different characteristic parameter sets were used by the HMM-based method, i.e. It's MFCC. Although the HMM and SVM methods show lower detection rates in terms of false positive rate(FP) and false negative rate(FN), their FP (0 percent) for HMM and FP (0.5 percent) is so poor that most pathological speech voice cases have been properly classified. The GMM-based approach offers superior levels of classification for train data and test data, according to the study. Nonetheless, compared to the uncertainty matrix, we can get small FN (0 percent) and FN (0.5 percent) respectively from

HMM-based and SVM-based approaches. The pathological speech voice possesses such distinctive characteristics that, according to its characteristic parameters, that classification method will recognize the data.

In 2007[33], Wenxi Chen.et.al proposed a classification methodology based on the support vector machine (SVM) to trace the various pathological voices. Sound signals were sampled from the pronunciation of a vowel "a" vocalized by 214 participants, including 181 patients suffering from various dysphonies (such as polypoid degeneration, spasmodic dysphonia adductor, vocal exhaustion, vocal tremor, vocal fold edema, hyper activity, and erythema) and 33 healthy subjects. Twenty-five acoustic parameters were determined for each object from the sampled data. Next, the original acoustic data set was converted into a new feature space using the main component analysis (PCA) process. A soft-margin SVM and three forms of kernels were tested to know the periphery of classification of stable and pathological voices. The findings were taken under review under various combinations of parameters and kernels. SVM-based approach appears to be very promising in pathological voice identification.

2.2.4 k-nearest neighbor (k-NN)

In 2011, John Labiak, et.al[34] avoided the long and heuristic process of training traditional Gaussian mixture-based models, the nearest neighbor-based techniques provide an approach to acoustic modeling This study is useful in solving the problem of selecting the distance metric for a phonetic frame classifier k-nearest neighbor (k-NN). Relative to two experienced Mahalanobis distances, the standard Euclidean distance is based on projections of nearest neighbors (LMNN) and locality preservation (LPP). To through the test time of classification k-NN, the locality sensitive hashing technique uses approximate nearest neighbor search. Comparison of the error rates of these approaches. Based on the task of phonetic frame classification of speech, the performance of baseline Gaussian mixture-based and multilayer perceptron classifiers compared. The classification of k-NN performs better than models of Gaussian mixture, but not multilayer perceptron. The proposed system finds the best performance in classification of k-NN using LPP, whereas LMNN is slightly less performance.

In 2010, k-NN / SASH phoneme classification algorithms were projected by Ladan Golipour.et.al[35] to compete satisfactorily with high-tech methods. The use of a parallel search algorithm (SASH) was effectively used to identify texts and photographs of high dimensions. Unlike other search algorithms, the data dimensionality does not affect SASH's computational time. Therefore, investigator uses their main frames and those of their margins to create fixed-length yet high-dimensional attribute vectors for phonemes. The phoneme classifier k-NN / SASH is fast, reliable, and could achieve a TIMIT test database classification rate of 79.2 percent. Finally, this algorithm's relevance is used to rescore phoneme lattices, created for context-independent and context-dependent tasks by the GMMHMM monophonic recognizer. In both cases, the identification of k-NN / SASH leads to changes in the level of recognition.

In 2012, two speech parameterization techniques LPCC and MFCC were used by Ooi Chia Ai et al.[36]. Comparisons were made for the available stuttered event data on the basis of repetitions and prolongation recognition. Experimental results showed that in all scenarios such as frame length selection, window overlapping percentage and a value in the first order high pass filter, LPCC gives better performance than MFCC. 21 LPCC features are 94.51 percent accurate. 25 MFCC features are 92.55 percent accurate. This is because LPCC is capable of capturing outstanding data from stuttered events, and the ability to distinguish all stuttered events, namely repetition and prolongation, is slightly increased. This study also reports that to classify repetition and prolongation, k-NN and LDA can be adopted as a classifier. Finally, to compare the accuracy of k-NN and LDA, the conventional validation was performed.

Li-Yu Hu, et.al. in 2016. [37] Proposed use for medical field data set of K-nearest neighbor (k-NN) classification. K-NN is a classifier that is not parametric. It has been used in many paradigm classification issues as the baseline classifier. To determine the final performance of the classification, it is based on calculating the distances between the test data and each of the training data. This research shows that the chosen distance variable affects the classification accuracy of the k-NN classifier. K-NN's Chi square distance function output is best performed on medical domain datasets including categorical, numerical, and mixed data types.

T.LaxmiPriya et.al. In 2012[38] a technique was suggested using certain features to increase robustness in a noisy environment. The K-NN method is used to identify these multiple features that are effectively combined. For the classification of each speech signal, speech and non-speech identification along with its classification method, which needs to be further enhanced in the detection of endpoints in noisy environments, is done. In speech communication, the diagnosis of vocal disorder is critical. Speaker uses subjective technique, but it's a tough process and can irritate patients. The aim of this research is therefore to provide early detection of vocal disorder. Dysarthria is a neurological disorder that causes damage to speech motor systems. The speaker has breaks in pitch, extreme pitch variability, excessive variations in loudness, changes in speech speed and prolonged intervals. This function is nullified by using speech processing software. To distinguish the difference between ordinary and dysarthric speech, the K-NN classifier is used. Investigation of different characteristics on classification of speech and non-speech in noisy environments was carried out. By using K-NN recognition algorithm accuracy, 80% K-NN classifier has been selected because it is less complex and easy to run.

K U Syaliman, et. in 2017[39] Al. Oh, al. suggested an approach to addressing majority voting problems by using a combination of local mean-based neighbor k-nearest (LMKNN) and distance weight k-nearest neighbor (DWKNN) to gain distance weight. The k-NN approach is one of the most commonly used data mining and machine learning analysis methodologies, such as categorization of text, pattern recognition, and classification. The k-NN is known as an attractive, easy-to-use, intuitive, less complex and can be used in different applications. It is found that the accuracy of k-NN is still relatively low compared to other classification algorithms. Many factors are responsible for the relatively low accuracy of k-NN. One of them is that in calculating distance, each characteristic of the method has the same result. The solution to this problem is to give characteristic weight to each data Another aspect that triggers the poor reliability of the k-NN is the selection of new data categories based on a simple majority voting system. Majority voting method neglects data closeness, which is not appropriate when the distance of each nearest neighbor varies considerably from the test data length. In addition, there will be the possibility of a

double majority class caused by the class determination system for new data based on the majority vote and the number of nearest neighbors where the number of nearest neighbors is chosen according to the desired level of success. Nonetheless, problems can be solved by using distance weight in deciding new data categories with a voting majority method which ignores the similarity between data, resulting in misclassification. The category determination of new data is based on the weights of the information length. So this work proposes a substitution of the voting majority model in k-NN using range weight approach to address the above disadvantages. Combine the k-nearest neighbor distance weight (DWKNN) and local k-nearest neighbor (LMKNN) methods to find the weight between data. The combination of these two methods improved the classification process resulting in accuracy. The precision of the tests is compared to the accuracy of the original k-NN system followed by several datasets. A result shows that it was possible to improve the classification accuracy of k-NN by combining LMKNN and DWKNN. The increase in accuracy is as high as 5.16 percent for the actual data. The highest score reaches 90.91% if k= 10 and the lowest is 84.85% if k= 1.

How K-Nearest Neighbor's main parameters influence his performance is discussed and his empirical evidence is presented by Gustavo E.A.P.A. Batista, et.al. In the year 2009[40]. The parameters investigated are the number of nearest neighbors, function of weighting, and function of distance. Evaluated the most common parameter options, including nine k values, three well-known weighting functions and three popular distance measurements. The Euclidean distance is commonly used and for qualitative attributes it is well adapted. One concern about this length, however, is that it does not handle inherently qualitative attributes of heterogeneous distance function using different attribute distance functions for different types of attributes is the best way to handle data sets with both qualitative and quantitative attributes. For example, one could use the overlap metric for qualitative attributes and quantitative attributes using the standardized Euclidean distance. This approach is called the Euclidean Overlap Metric (HEOM) heterogeneous approach. Instead of the Euclidean distance defined as Heterogeneous Manhattan-Overlap Metric (HMOM), another approach uses Manhattan distance metric. One criticism of the overlap metric is that additional information on qualitative attributes is not made available. The Value Difference

Metric (VDM) is an approach to overcoming this barrier. The classification similarity is used in the VDM method to measure the differences between these values for each possible value of an element. As a result, a distance matrix is generated for every attribute from the training set. For each value, the statistical sample is unreliably small and while all numerical attribute values are not unique, there are still so many random values and the calculation of distance is still untrustworthy. It is not advisable to use the VDM directly on quantitative attributes because of these problems. Discretization approach can solve the problem of using VDM on qualitative attributes. It is possible to discrete a quantitative attribute and treat it as a qualitative attribute. Discretization will result in the loss of significant quantitative information available. The Heterogeneous Distance Function (HVDM) similar to HEOM, except that it uses VDM for qualitative attributes instead of an overlap metric and standardizes differently as well. This research describes how the parameters affect the behavior of the algorithm k-Nearest Neighbor. This research recommends the use of the inverse weighting function as a result of the results obtained. This weighting function has two major advantages: it has excellent mean performance, outperforming statistically the other two weighting functions; and it causes the parameter k to have a smooth influence on the algorithm's classification performance, as it penalizes distant neighbors. This research emphasizes that between 5 and 11 nearest neighbors are obtained the highest accuracy results for k. The research recommends that the inverse weighting function be used, k = 5 showed the best mean performance for HEOM and HMOM, and k = 11 showed the best mean performance for HVDM. Furthermore, the test could not notice any significant differences between the effects of the three distance functions: HEOM, HMOM and HVDM. Since HVDM can only lead to classification in the presence of qualitative attributes, the subset of data sets with at least one qualitative attribute has been chosen.

2.3 Research Gap and Challenge

Literature survey shows that different techniques were proposed to understand the dialogue of people with articulatory disabilities. There are few drawbacks in each of these techniques. And so none of the methodology was fully developed yet. Existing methods developed and suggested methods are not efficient and effective by researchers for speech recognition of articulatory handicapped persons and therefore

error created by the device is high compared to the human system. There are various speech recognition systems for people with articulatory disabilities, but less attention is given to speech rectification. There is no standard articulatory disability database available that creates the major research issue. Database quality variations depend on the speaker. Because of this, generalizing the system in every context is a major challenge. Finding ways to bridge such a performance gap is now increasing interest. An effective way to represent, store and retrieve the knowledge and information required for natural conversation and minimize the error is needed. While substantial progress has been made over the past few decades, there is wide scope for the further improvement of the system for people with articulatory disabilities and the use of technology to benefit society.

Some of the major challenges that require more sophisticated procedures to manage it are listed on the basis of the literature review performed. The challenges are listed here and are divided into two groups.

While recent advances in automatic speech recognition have been made, robust and accurate speech recognition is still a challenging issue due to complex factors such as speech and content variations and distortion of the environment. Overcoming the speaker's variability is a major challenge in speech recognition. Speech phrases or continuous changes in dialog based on the speaker's emotional state and are also affected by the speaker's individual characteristics. Therefore, speech in the speaker independent mode is recognized primarily from the voice; this accomplishment of an independent speaker system is a breakthrough. A practical challenge is to consider expressions of emotion in the recognition of speech, and the system should also take explicit account of gender variability. The latest methods are also not tested against diverse noise and reverberation conditions in real-life applications.

One of the main research problems in speech recognition is how discriminative, affect-salient features can be derived from speech signals. Several features of speech are mentioned in the literature. Therefore, the challenge here is to include the four categories correctly:

1) Acoustic features,

2) Linguistic features (words and discourse),

3) Contextual information (e.g., topic, gender and turn-level features representing local and global aspects of dialog), and

4) Hybrid features combining acoustic features with other information.

Two major issues need to be handled for building recognition systems. They are

i) Selection of right features to comprise the unique information to be easily recognizable by any classification model.

ii) The right selection of samples for training a classification model.

The methods of speech analysis need to handle the following characteristics of the signal that directly affect the system of speech recognition. (III) Great variability of utterances that can be pronounced and no evidence of the best choices, (IV) considering the discrepancy between the speech characteristics of the interspeaker and the optimum classification model training. A traditional feature called MFCC has recently been used to recognize speech. If the noisier data is presented, MFCC fails to extract significant feature values from the speech signal.

It is very challenging to develop a noise-adaptive classification algorithm for voice recognition. Most speech processing methods in the literature use HMM and GMM to classify emotion, but the main problem with these methods is that they require detailed assumptions about the parameters of the data distribution and the model. GMM has been used for classification that is not suitable for noisy environmental conditions. GMM also requires more data samples to be used for consistent training. Neural network-based classification models also require a lot of training data for better classification, and a low-level feature, local order; these acoustic models are hard to handle with intrinsic characteristics.

Work is limited to improving the most significant of these is the need to improve the accuracy of existing speech recognizers before being able to incorporate them with confidence into the lives of the disabled community.

2.4 Summary

The section addressed the different types of speech impairments and methods for speech recognition. Researchers focused on various feature extraction techniques such as autocorrelation function and envelope parameters, duration and frequency of dysfluent portions, speaking frequency, frequencies of 1st to 3rd formant and

amplitude Mel Frequency Cepstral Coefficients (MFCC) spectral measurement. After using MFCC feature extraction technique as this is perceptually driven, most researchers obtained the best results. Some of the researchers have tried various techniques of removal of features in combination. HMM, ANN, SVM and KNN are among the most common classifiers chosen by the researchers. Classifier output depends on the choice of apps. Results reliability can be achieved through intervals of cross-validation and trust. Cross validation is a technique used to estimate a classifier's accuracy. Studies focused on speech recognition in people with articulatory disorders, but not on disability speech rectification. This gives motivation in this area to work. The next chapter addresses the challenges of developing speech recognition and rectification algorithms for people with articulatory disabilities.

Chapter 3

PHYSIOLOGY OF SPEECH PRODUCTION MECHANISM AND PROCESSING OF SPEECH RECOGNITION

3.1 Overview

Speech is the easy way to communicate with the other people. All are learning the appropriate skills during early childhood. The phenomenon of speech production is very complex but it comes very naturally for the normal person. Human vocal tracts and articulators are biological organs with nonlinear properties, the activity of which are not only under conscious control but also influenced by factors ranging from sex to upbringing to emotional state.

As a result, vocalizations may vary widely in terms of accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed; furthermore, during transmission, our irregular speech patterns may be further distorted by background noise and echoes, as well as by electrical characteristics. All these variance sources make speech recognition a very complex issue, even more than speech generation.

Speech recognition is a process used to interpret speech stimuli into understandable words or combinations of understandable words. Speech recognition is the area in which speech signals are analyzed and how these signals are processed.

Humans use their vocal cords to produce a quantity of sound signal and for recording a sound a high quality microphone is used. After that, the speech recognition system is used to recognize this signal and convert it into a series of words. This Chapter presents a broad description of how the speech is produced and it also explains types of speech disorder and their consequences in speech recognition system.

3.2 Fundamentals of speech production Mechanism

To ordinary people, speech is just the sound waves that come from the human mouth and are perceived / heard through the ears. But behind its production there is a complex mechanism. Studying the production and understanding of human speech is important and appropriate for the creation of hearing aids, cochlear implants, speech recognition, speech enhancement, speech simulation, speech modeling etc. The entire voice production mechanism consists mainly of three functions represented in Figure 3.1 [1] by block diagram. Motor control is the human brain-driven function that

generates a thought of what to speak and therefore provides the organs of speech production with control signals through sensory nerves. When the command signals are received from the motor control unit, talk produced. The whole mechanism, referred to as the' Articulatory Motion'. The next function of the human speech production mechanism is the speech generation, which consists of the air that comes out of the mouth and the nasal cavity and is thrown into the open space in the form of an acoustic wave. In case of speech perception, an acoustic wave generated by the mouth and the nasal cavity reaches the human ear and is perceived by the sensory nerves that connect the ear with the human brain.

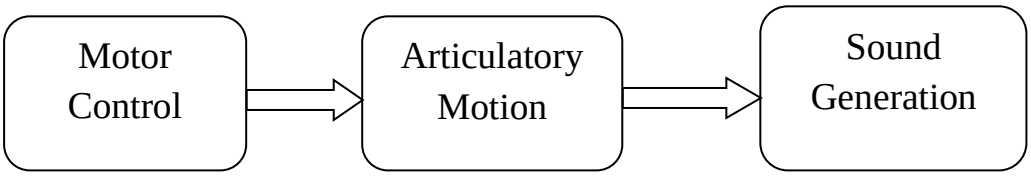


Figure 3.1: Block diagram of the development process of human speech

Various people's speech varies in their basic parameters such as fundamental frequency, range, and pitch. It is quite difficult to produce the exact replica of one's speech through machine simulation. Figure 3.2 gives the idea about human brain segmentation.

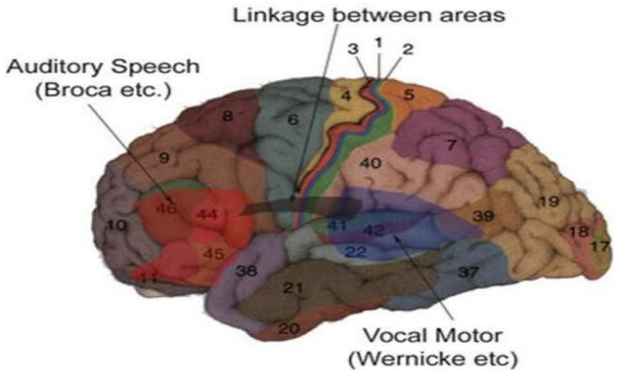


Figure.3.2: The human brain segmentation displaying regions of Brodmann, (Source : 2014, Harish Chander Mahendru)

3.2.1 Motor Control Function

Humans almost naturally speaking to each other without even knowing what will happen on behind the mechanism of speech production. Generation of idea in the mind about what to talk at that moment the production of speech starts. The generation of thoughts is passed on through sensory nerves to the human vocal apparatus. The complete phenomenon is called motor control. The language

processing and the motor commands generation are the two parts of motor control [2]. Our brain is divided into different segments. These segments are responsible to perform various control, think and memory functions [2]. This is shown in Figure.3.2 the human brain segmentation displaying regions suggested by Brodmann [3]. The scientist will be curious to know the answers to questions like: How does the brain derive speech from an acoustic sound? How is voice sounds distinguished from background noise by the brain? How is a particular language's phonological structure perceived in the brain? How does the brain store words that a person knows and access them? How does the mind turn terms into constituents and phrases? What is the use of structural and linguistic knowledge to understand sentences? Phonetics, phonology, morphology, lexicology, syntax and semantics are specific neurolinguistics subfields through which researchers have discovered the answers to the above-mentioned questions [6], [7]. Within Figure 3.2, the area is the auditory zone, shaded red, numbered 44, and the area is shaded blue, numbered 41 & 42 is the region of motor control. In reality, through other sensory organs of language processing, the auditory region (Broca's region) receives feedback in the form of listening and visual signals, helping it to determine what to say or what sound to produce [8]. Accordingly, the motor control area (the region of Wernicke) produces control signals to transfer the vocal tract apparatus and other organs of speech development such as lungs, vocal cords, glottis, jaw, tongue, teeth, lips etc. The entire process is shown with Figure 3.3.

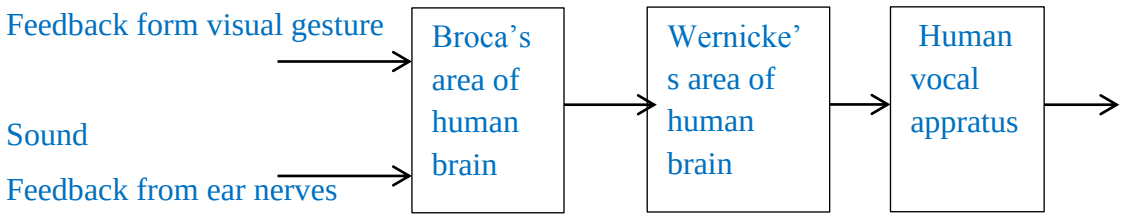


Figure 3.3: Sound generation and production process, (Source: 2014,Harish Chander Mahendru)

3.2.2 Articulatory Motion

Different organs are involved in human speech and sound development. These organs are versatile in nature and adjust their shape and size according to the motor control signals obtained from the brain, depending on the type of speech or sound to be made. Lungs provide the required air force in the form of an acoustic wave to produce noise. The air passes through the vocal tract, vocal cords, glottis, epiglottis, and other organs

in the mouth and eventually comes out in the form of an acoustic wave through the mouth and nasal cavities. Different organs that the air passes through during the speech and sound generation process are shown in the Figure 3.4 [9],[10].

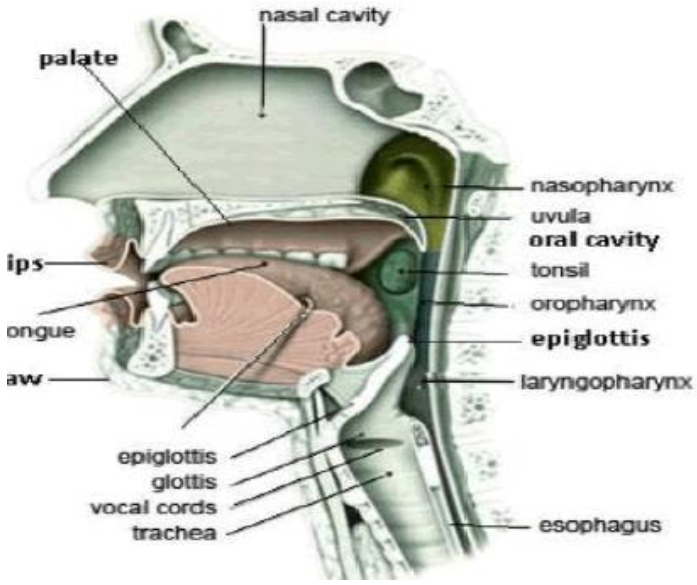


Figure 3.4: Structure of Vocal Apparatus, (Source: Matt Edwards)

The air expelled from the lungs moves through the trachea while speaking and enters into the larynx. A pair of lip-like tissues called vocal cords regulates the air in the larynx. These are very important vocal apparatus membranes who decide the pitch of the emitted voice. The vocal cords are pearly white in color fixed at one end to the cartilage of the arytenoids on the back and to the cartilage of the thyroid on the front. Compared to the high pitch voice for females with small larynx, males generally have low pitch voice with large larynx.

The width of male vocal folds ranges between 17 and 25 mm and varies between 12.5 and 17.5 mm for females. The vocal folds vibrate to produce articulated speech and temporarily limit unvoiced speech. In addition, there are different types of speeches called phonemes for which vocal cords open and close in different ways to allow the air to pass through them and bring them to the top. Vocal tract is the passage-like tube running from glottis at one end and at the other end with two openings, oral and nasal cavity. It is a non-uniform cross section with an approximate length of about 17 cm for males. It branches out at the soft palate (velum), just halfway through the tract, and opens up at the nose as a second branch. This part of the vocal tract is approximately 13 cm long. Air, after leaving the vocal cords, enters the pharyngeal,

mouth and nasal cavities which, by amplifying some of the frequencies and attenuating others, provide the required resonance to the sound as per term. Other organs in the mouth, such as soft palate, teeth, tongue, lips, jaw, alter their shape and move accordingly to block or allow the air to escape the mouth and nose, thereby modulating the sound to give required form and amplitude. Each individual's speech is distinctive due to the difference in size and shape of different speech output organs. The uniqueness of the entire articulatory movement system is that it can respond very quickly to the rapidly changing speech parameters even after having so many complexities. Under the pharynx, epiglottis and false vocal cords play an important role in preventing food from reaching the larynx and acoustically isolating the esophagus from the vocal tract.

3.2.3 Phonemes

People generally speak the language depending on the region in which they are brought up. To speak in their mother tongue, one does not need special training or knowledge. Children learn to speak with the audio and visual movements at an early age of one year. With the aid of symbols called phonemes, the word signs of any language can be pronounced. Set of phonemes are used to express all the words with different tones of any language. There are 20 to 60 phonemes in all the languages spoken in the world [11], [12].

Phonemes of any language include the situational impact, emotions and characteristics of the speaker to be pronounced which is, of course, not necessary for the language's written text. Phonemes are designed on the basis of articulatory movement of the vocal tract. English language consists of 40 phonemes [13].Vowels and consonants are the two parts of phonemes. Voiced part of sound is always vowels but consonants may be voiced or unvoiced. When air passes through vocal cord, it vibrates periodically with fundamental frequency about 110 Hz for men, 200 Hz for women and 300 Hz for children. These periodical vibrations are nothing but voice part of sound. Besides the basic level, the articulatory motion of the organs of speech production produces resonance frequencies according to the phoneme. The resonance frequencies "N" number, F_1, F_2, ..., F_n are called Formant Frequencies. F1 = 180-800 Hz, F2 = 600-2500 Hz, F3 = 1200-3500 Hz, and F4 = 2300-4000 Hz are the normal range of formant frequencies for adult males. On the other hand, the unvoiced sound

in nature is completely random. The vocal cords are completely open, completely close or partially open during the development of unvoiced speech. Vowel phonemes are created by repeated vocal cord vibration. Vocal phonemes are further classified into three groups according to the position of the tongue in the mouth cavity, namely "Front" such as /IY/, /IH/, /EY/, & /EH/; "Mid" such as /AA/ & /ER/ and "Back" such as /AE/, /AO/, /UH/, /OW/, & /AH/. Diphthong sounds are produced in a single syllable as /AY/, /AW/, /OY/ & /UW/ by switching between two vowels. Semivowels are formed when the vocal tract is completely closed by the tongue or lips [].

There are two types of semivowels, "Glides" such as /Y/ & /R/ and "Liquids" such as / W/ & /L/. Consonants can be voiced or unvoiced and marked as Nasal, Stop and Plosives, Fricative or Affricate. Nasal Fast Analysis of Human Speech Production Mechanism 53 sounds are made when the mouth cavity is closed and the air moves through the nasal cavity through the opened velum. Plosive sounds like /P/, /B/, /T/, /K/, /D/ & /G/ are created when, due to its momentary closing, pressure built up behind the vocal cords is suddenly released. Here /B/,/D/&/G/ is voiced and /B/,/T/ & /K/ is not voiced. If the mouth cavity is not completely blocked and the air flow is quasi-periodic due to vocal cord vibrations, the sound produced is named as fricative sounds such as /HH/, /F/,/V/, /TH/, /DH/,/S/,/Z/, /SH/ & /ZH/. Affricate sounds like/CH/ & /JH/ are produced by double plosive action followed by fricative action.

3.2.4 Concept of human speech production mechanism

The figure.3.5 shows the mechanism of human speech production and following points describes the procedure of human speech production.

(i) Speech signal is produced when the vocal cords producing sequence sounds are acoustically excited by the air expelled from the lungs.

(ii) The lungs together with the diaphragm are the main source of speech signal output. (iii) There are three main vocal tract cavities:

(1)Pharynx (2) Oral cavity (3) Nasal cavity

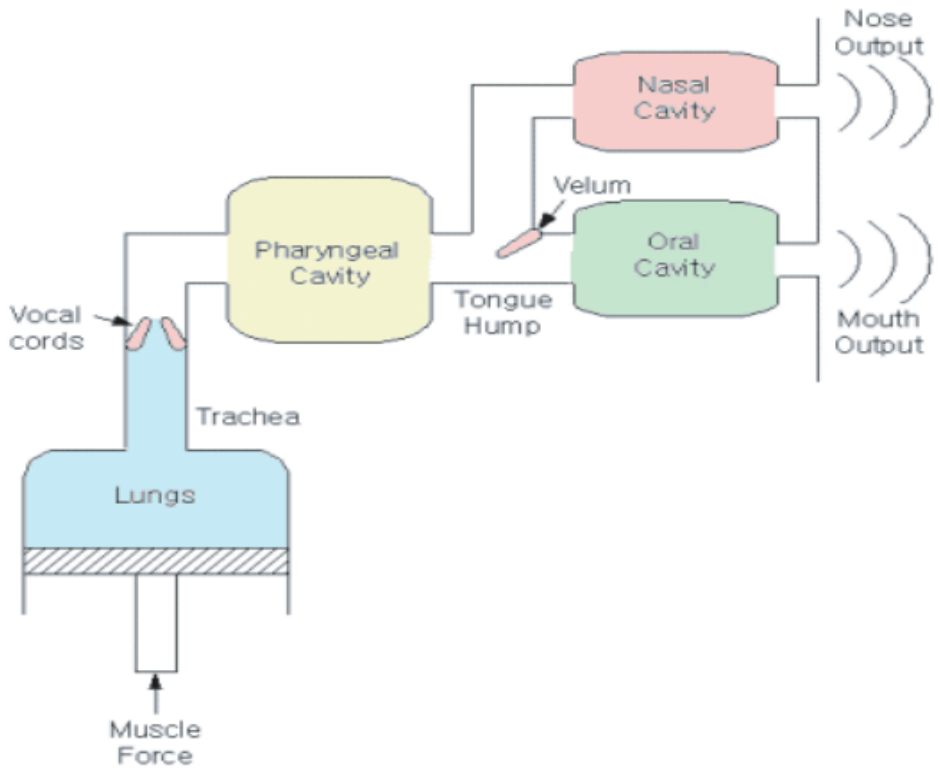


Figure 3.5: Mechanism of human speech production (Source: Laura Docio-Fernandez)

(iv) The air flows through the V-shaped opening called the glottis and the larynx into the vocal tract.

(v) The basic purpose is to control the airflow by rapidly opening and closing the valves in the vocal cord, producing a variety of sounds.

(vi) The frequency of vibration depends on mass and tension. It varies from individual to individual.

(vii) Smooth palate from the nasal cavity to the pharynx acts as a device to interact and isolate them from each other. The extreme end of the pharynx consists of epiglottis and false vocal cord that helps to prevent the food from entering the larynx and is closed during chewing while opening during breathing.

(viii) By moving our lips, tongue, palate, teeth and cheeks, the acoustics can vary. It depends on its size and shape as well. Sound is also generated by the walls and construction in the vocal tract.

(ix) Lung, larynx, and vocal tract are the organs taking responsibility for speech, in which the lungs provide the airflow to that of the larynx, while the larynx then alters it to produce noisy vocal tract air flow.

(x) The basic types of sound are intermittent, disruptive and impulsive sources, where they are usually used in combinations. Example: the word 'stop' uses all three sources where, |st| is disruptive, |o| intermittent and |p| impulsive.

3.3 Speech Signal Representation

Speech signal consists mainly of three parts, which are often voiced, unvoiced and silent. The voice portion is of a periodic nature and the unvoiced portion looks like a random signal. These two kinds of speech signals are shown in Fig.5. Speech signals are represented in various forms such as time domain representation, frequency domain representation and spectrogram. These forms of representations can be used for speech analysis. These three types of representations are given in Figure 3.5 (a, b, c, d).

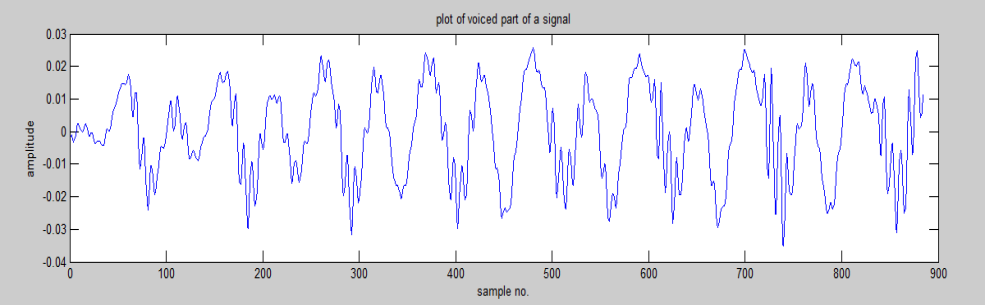


Figure 3.5(a): Plot of voiced part of vowel "a"

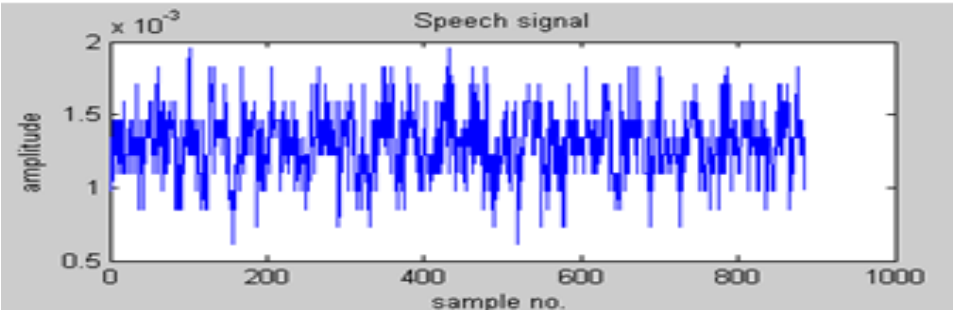


Figure 3.5(b): Plot of unvoiced part of vowel "a"

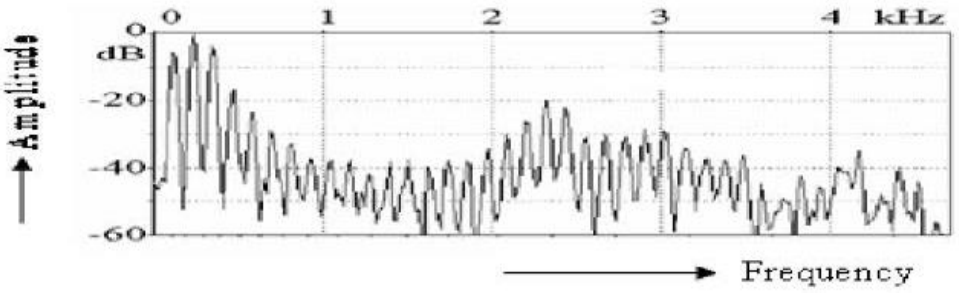


Figure 3.5(c): Plot of Frequency verses Amplitude for speech signal (Source: Fant)

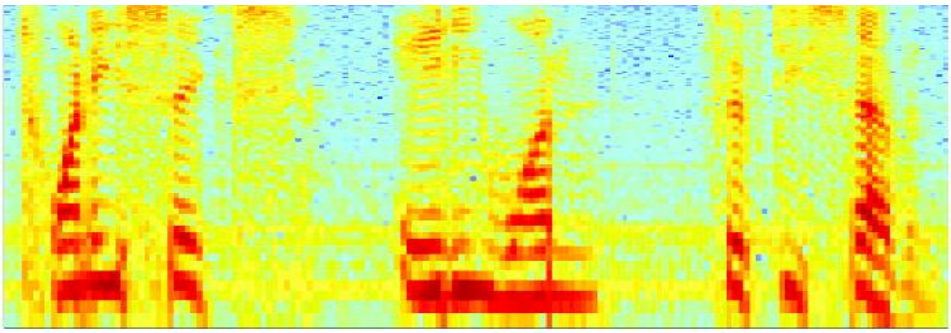


Figure 3.6(d): Spectrogram of speech signal (Time verses Frequency) (Source: Muhammad S A Zilany)

Time vs. amplitude is interpreted in the time domain representation signal. Frequency vs. amplitude is described in the frequency domain representation signal. And the signal is interpreted as time versus frequency in spectrogram representation. From the point of view of speech synthesis and analysis, all these representations are essential. Due to its very non-stationary behavior of parameters (that is changing at every 50 – 100 ms) such as pitch and loudness, the study of speech signal is again very tedious. For short time the nature of speech signal is expected to be stationary and so for the speech analysis and synthesis short time analysis is mostly used. Windowing technique is used for short time analysis. Different methods are used for this analysis such as short time zero crossing rate, short time energy, short time auto correlation function, short time Fourier transform, short time z-transform, short time cepstrum, short time homographic filtering and short time spectrograph.

3.4 Human Attributes

Physical attributes are low-level signal features that identify key aspects of the signal's spatial or spectral properties**.** Although some of the features are perceptually driven, they are known as physical features as they are directly determined from the amplitudes of the audio waveform or the corresponding short-term spectral values. In the following equations, the "a" sub index indicates the current frame so that $x_a[n]$ are

samples of the N-length data segment (possibly multiplied by a window function) corresponding to the current frame. The analysis of the "a" th frame is

$$\left\{ \begin{matrix} x_a[n] \\ n = 1 \ldots N \end{matrix} \right\} \rightarrow \left\{ \begin{matrix} x_a[n] \text{ at freq. } f[k] \\ k = 1 \ldots N \end{matrix} \right\} \quad (3.1)$$

3.4.1 Zero-crossing rate

The Zero-Crossing Rate (ZCR) measures the number of times the sign changes in the signal waveform during the current frame and is indicated by

$$ZCR_r = \frac{1}{2}\sum_{n=1}^{N}\left|sign\left(x_r(n)\right) - sign\left(x_{r-1}(n)\right)\right| \quad (3.2)$$

Where,

$$sign(x) = \begin{cases} -1, & x < 0 \\ 1, & x \geq 0 \end{cases} \quad (3.3)$$

The ZCR generates spectral data at a low cost for several applications. The ZCR is directly related to the basic frequency of narrowband signals (e.g. a sinusoid). The ZCR is well associated with the average frequency of the main concentration of energy for more complex signals. The short-time ZCR takes on values for speech signals that fluctuate rapidly between voiced and unvoiced segments due to their varying concentrations of spectral energy. On the other hand, the ZCR is more stable over extended time periods for music signals.

3.4.2 Short-Time Energy

It is the mean square value in the waveform values data frame which reflects the temporal boundary of the signal. The variation over time can be more than its actual magnitude a good predictor of underlying signal value. It will be measured as,

$$E_a = \frac{1}{N}\sum_{n=1}^{N}|x_a(n)|^2 \quad (3.4)$$

3.4.3 Autocorrelation

$$R_n(k) = \sum_m x_n(m)x_n(m-k) \quad (3.5)$$

$R_n(k)$ is a measure of how similar $x(n)$ is to $x(n-k)$, so it is useful for pitch detection.

Even function, $R_n(k) = R_n(-k)$.

$R_n(k)$ is the projection of $x_n(m)$, onto $x_n(m-k)$, so it is maximum at $k = 0$:

$$\sum_{m=k}^{L-1} x_n(m)x_n(m-k) \leq \sqrt{\sum_{m=k}^{L-1} x_n(m)^2} \sqrt{\sum_{m=k}^{L-1} x_n(m-k)^2} \leq \sum_{m=k}^{L-1} x_n(m)^2 \quad (3.6)$$

If $x(n)$ is periodic, $R_n(k)$ approximately periodic with same period:

$$x_n(m) = x_n(m+N) \quad (3.7)$$

$$x_n(m-N) = x_n(m) \quad (3.8)$$

$$R_n(k+N) = \sum_m x_n(m)x_n(m-k-N) \quad (3.9)$$

$$\approx \sum_m x_n(m)x_n(m-k) \quad (3.10)$$

$$= R_n(k) \quad (3.11)$$

3.4.4 Band-level Energy

This refers to the energy within the specified signal spectrum frequency range. It is possible to calculate the corresponding weighted sum of the power spectrum as indicated by

$$E_r = \frac{1}{N}\sum_{k=1}^{\frac{N}{2}} (X_r[k]W[k])^2 \quad (3.12)$$

W[k] is a weighting feature of non-zero values over only the finite range of the bin indices "k" which corresponds to the frequency band of concern. Sudden changes in band-level energy indicate a change in the distribution of spectral power, or timbre, of the signal, which aid in the segmentation of sound. Energy log transformations are commonly used to increase spread and reflect relative (perceptually more relevant) differences.

3.4.5 Spectral-centroid

It's the degree continuum's center of gravity. It is a spectral structure gross metric. If the content of high frequency is lower, the location of the centroid spectral frequency is high.

$$C_r = \frac{\sum_{k=1}^{\frac{N}{2}} f[k]|X_r[k]|}{\sum_{k=1}^{\frac{N}{2}} |X_r[k]|} \qquad (3.13)$$

Since shifting the major energy density of a signal to higher frequencies makes it sound clearer, the spectral centroid has a strong correlation to the subjective perception of sound intensity.

3.4.6 Fundamental Frequency (F_0)

It is determined by calculating the time-domain waveform periodicity. It can also be determined from the sound continuum as the amplitude of the first harmonic or as the distance between the harmonics of the periodic wave. For human voice, the approximation of F0 is a non-trivial problem owing to (i) the period-to-period variance of the waveform, and (ii) the fundamental frequency part may be low compared to the other harmonics. This causes the duration captured to be prone to doubling and halving errors. Time-domain periodicity can be estimated from the autocorrelation function of the signal given by,

$$R(\tau) = \frac{1}{N}\sum_{n=0}^{N=1}(x_r[n]x_r[n-\tau]) \qquad (3.14)$$

Where, $R(\tau)$ indicate local maxima at the pitch period and its multiples. The inverse of the lag "τ" estimates the fundamental frequency that corresponds to the maximum of $R(\tau)$ within a specified range. Considering short lags over longer ones, fundamental frequency multiples is avoided. The normalized value of the lag at the expected period signifies the strength of the signal periodicity and is mentioned to as the harmonicas coefficient

3.4.7 Mel-Frequency Cepstral Coefficients (MFCC)

MFCC are perceptually motivated features which provide a compact representation of the envelope of the short-time spectrum. Most popularly used feature set in speech recognition and in music also [11]. To calculate the MFCC, a DFT transforms the windowed audio data frame. Next, in the frequency domain, a Mel-scale filterbank is applied and the power within each sub-band is calculated by squaring and summing the spectral magnitudes within bands. The Mel-frequency scale is linear below 1 kHz

and logarithmic above this frequency, a perceptual measure such as the critical band scale .Eventually, by applying a DCT to obtain the cepstral coefficients, the logarithm of the bandwise power values is taken and decorrelated. The log transformation helps to deconvolve multiplicative spectrum components such as the function to pass the origin and filter. The decorrelation results in the concentration of most energy in a few cepstral coefficients. For example, in 16 kHz of speech sampled, 13 low-order MFCCs are sufficient to represent spectral envelopes across phonemes. As the difference between the signal spectrum and the spectrum reconstructed from the popular low-order cepstral coefficients, a related feature is the cepstral residual measure.

3.4.8 Spectral Roll-off

It is another most commonly used attribute and it is represented as,

$$R_r = f[k] \tag{3.15}$$

Where, K is the largest bin that fulfills

$$\sum_{k=1}^{K} |X_r[k]| \leq 0.85 \sum_{k=1}^{\frac{N}{2}} |X_r[k]| \tag{3.16}$$

The roll-off is nothing but the frequency below which 85% of accumulated spectral magnitude is concentrated. It takes on higher values for right-skewed spectra.

3.4.9 Spectral Flux

The frame-to-frame square difference of the spectral magnitude vector is summarized as,

$$F_r = \sum_{k=1}^{\frac{N}{2}} (|X_r[k]| - |X_{r-1}[k]|)^2 \tag{3.17}$$

This gives an indicator of the change rate of the local spectral. A high spectral flux value suggests a sudden shift in spectral magnitudes and thus a potential segment boundary in the r^{th} frame.

All of the features listed above so far were short-term parameters measured at frame rate from windowed length audio segments not exceeding 40 ms, the presumed period

of stationary audio signal. The temporal pattern of changing signal properties observed over a sufficiently long interval is an equally important predictor of signal identity.

3.5 Perceptual Features

Human sound perception is based on the sound's sensory attributes. If there is no good source template, perceptual characteristics provide an alternate basis for segmentation and classification. Psychic sensations evoked by sound can be broadly categorized as loudness, pitch and timbre. Loudness and pitch can be ordered from low to high on a magnitude scale. On the other hand, Timbre is a more synthetic, multi-dimensional experience that distinguishes different sounds of the same loudness and pitch. To obtain numerical representations of short-time perceptual parameters from the audio waveform segment, an auditory system computational model is used. Loudness and pitch are common perceptual features along with their temporal fluctuations and are briefly reviewed as follows:

- **Loudness**

Loudness is a representation of the strength of the signal. It is associated with the sound intensity as predicted, but it also depends on the sound length and range. The perceived loudness is measured in physiological terms by the total sum of the auditory neural activity produced by the noise. Loudness varies with sound frequency nonlinearly. Accordingly, loudness computation models achieve loudness by summing critical band filter contributions to a compressive power [12]. Significant aspects of loudness perception captured by loudness models are nonlinear loudness scaling with intensity, loudness frequency dependence and loudness additivity across spectrally separated components.

- **Pitch**

While pitch is a perceptual attribute, it is closely related to the basic frequency (F0) physical attribute. Changes in subjective pitch are related to the F0 logarithm so a constant pitch change in music refers to a constant ratio of fundamental frequencies. Most of the pitch detection algorithms (PDAs) derive F0 from the acoustic signal, i.e. they are focused on the calculation of the periodicity of the signal through the repetition frequency of different temporal features, or on the analysis of the harmonic

structure of the continuum. The auditively inspired PDAs use the cochlear filter bank to decompose the signal and then measure the periodicity of each channel independently via the ACF [13]. Due to the massive higher channel bandwidths in the high frequency region, multiple lower harmonics are mixed in the same channel and the observed periodicity corresponds to that of the baseline frequency amplitude envelope beating. The perceptual PDAs seek to mimic the robustness of the ear to interference-corrupted signals as well as to generate slightly harmonic signals that still provide a good pitch sensation. Modulation energies in the 20-40 Hz range of bandpass filtered audio signals are associated with the perceived roughness of the tone, whereas modulation energies in the 3-15 Hz range are representative of speech syllabic rates [8].

3.6 Different short time analysis used in speech recognition

3.6.1 Short –time analysis

Speech is non-stationary, but all of our analytical tools assume a stationary signal. So the first step in speech analysis is to cut it into \frames, which are short enough (with 20ms) to be considered stationary. Define as follows

$$x_n(m) = x(n-m)w(m) \qquad (3.18)$$

Where, $w(m)$ is a finite-length windowing function of length L; for example, the rectangular window $w(m) = u(m) - u(m-L)$. The window will move between frames by M points and show a matrix where each frame is a matrix row

$$:\begin{bmatrix} x(0) \\ x(1) \\ \cdots \\ x(n) \\ \cdots \end{bmatrix} \rightarrow \text{Convert to Frames} \rightarrow \begin{bmatrix} x_0(0) & \ldots & x_0(m) & \ldots & x_0(L-1) \\ x_M(0) & \ldots & x_M(m) & \ldots & x_M(L-1) \\ & & \ldots & & \\ x_n(0) & \ldots & x_n(m) & \ldots & x_M(L-1) \\ & & \ldots & & \end{bmatrix} \qquad (3.19)$$

Frame rate: there is one frame every M samples, or F_s/M frames/sec.

Overlap: neighboring frames overlap by $L - M$ samples.

3.6.2 Short-Time Fourier Transform

The short-term transformation of Fourier is a function of both the n and ω:

$$S_n(e^{j\omega}) = \sum_{m=0}^{L-1} s(m)w(n-m)e^{-j\omega m} \qquad (3.20)$$

Two interpretations:

1. n fixed – Fourier transform:

$$X_n(e^{j\omega}) = \mathcal{F}\{x(m)w(n-m)\} \quad (3.21)$$

2. ω fixed – Frequency-shifted bandpass filter:

$$X_n(e^{j\omega}) = x(n)e^{-j\omega m} * w(n) \quad (3.22)$$

3.6.2.1 Fourier Transform Interpretation

Fourier transforms definition STFT is like a window spectrum-converted

$$\text{DFT:}S_n(e^{j\omega}) = \mathcal{F}\{s(n)w(n-m)\} \quad (3.23)$$

$$= \mathcal{F}\{s(n)\} * \mathcal{F}\{w(n-m)\} \quad (3.24)$$

$$= S_n(e^{j\omega}) * W(e^{-j\omega})e^{-j\omega} \quad (3.25)$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} S(e^{j(\omega-\theta)})W(e^{j\theta})e^{-j\theta}d\theta \quad (3.26)$$

3.6.2.2 Filter bank Interpretation

$$S_n(e^{j\omega_i}) = e^{j\omega_i n}\sum_{m=0}^{L-1} s(m)w(n-m)e^{j\omega_i(n-m)} \quad (3.27)$$

$$= e^{-j\omega_i n}[s(n) * w(n)e^{j\omega_i n} \quad (3.28)$$

$$= e^{-j\omega_i n}s_i(n) \quad (3.29)$$

Where, the filter bank $h_i(n)$ was defined as follows:

$$s_i(n) = s(n) * h_i(n), \quad h_i(n) = w(n)e^{j\omega_i n} \quad (3.30)$$

$w(n)$ can be any FIR filter lowpass. It may be a rectangular window, or a Hamming window, or it may be a FIR approximation to an ideal LPF, developed with matlab t echnology. Suppose w (n) has the following features:

The window length is L: $w(n)$ is non-zero for $0 \le n \le L-1$.

The filter bandwidth is $2\omega_c$: $A(e^{j\omega})$ is non-zero for $|\omega| < \omega_c$, and approximately zero for all otherω.

3.6.3 Implementing non-uniform filter banks using the STFT

The frequency distribution of the filters is non-uniform. Non-uniformly spaced filters can be obtained by simply adding neighboring channels from a filterbank based on STFT. For example:

$$x_i(n) = \frac{1}{2}\left(s_i(n) + s_{i+1}(n)\right) \quad (3.31)$$

$$= s(n) * \frac{1}{2}\left(w(n)e^{j\frac{2\pi in}{N}} + w(n)e^{j\frac{2\pi(i+1)n}{N}}\right) \quad (3.32)$$

$$= s(n) * w(n)\cos\left(\frac{\pi n}{N}\right)e^{j\frac{2\pi n}{N}\left(i+\frac{1}{2}\right)} \quad (3.33)$$

This leads to a new filter with the following features:

Bandwidth: $2\omega_c + \frac{2\pi}{N}$

Center frequency: $\frac{2\pi n}{N}\left(i + \frac{1}{2}\right)$

Filter length: L

3.7 Window Characteristics

Different windows of analysis $\mathbf{w(n)}$ can produce different results of analysis. STFT, for instance, is like a DFT converted to the $\mathbf{W(e^{jw})}$. In particular, consider that $\mathbf{W(e^{jw})}$ to look as much as possible like an impulse ($\delta(\omega)$) In particular, for proper filtering As small as possible, the central lobe. As low as possible, side lobes.

It is not possible to obtain both narrow main lobe and low-amplitude side lobes, so it is necessary to determine a system-based technology trade-off requirement.

3.7.1 Hamming and Hanning Windows

$$w_2(n) = w_1(n)(\beta_1 - 2\beta_2\cos(\alpha n)), \quad \alpha = \frac{2\pi}{L-1} \quad (3.34)$$

$$W_2\left(e^{j\omega}\right) = \beta_1 W_1\left(e^{j\omega}\right) - \beta_2 W_1\left(e^{j(\omega+\alpha)}\right) - \beta_2 W_1\left(e^{j(\omega-\alpha)}\right) \quad (3.35)$$

Separating exponential terms, and by simplifying, the below equation is as follows:

$$= e^{j\frac{\omega(L-1)}{2}}\left[\beta_1\frac{\sin\frac{\omega L}{2}}{\sin\frac{\omega}{2}} + \beta_2\frac{\sin\frac{(\omega+\alpha)L}{2}}{\sin\frac{(\omega+\alpha)}{2}} + \beta_2\frac{\sin\frac{(\omega-\alpha)L}{2}}{\sin\frac{(\omega-\alpha)}{2}}\right] \quad (3.36)$$

For each sidelobe, the first term in the sum is 180^0out of phase with the second and thi rdterms, so it is possible to arbitrarily close the sidelobe of $W_2(e^{j\omega})$
to zero by selecting the correct 1 and 2. Some typical choices:
Hanning window:$\beta_1 = 0.5, \beta_{2=}0.25$
Hamming window: $\beta_1 = 0.54, \beta_{2=}0.23$

The first side lobe amplitude is less than 1 percent of the main lobe amplitude with these constants-that is, the first side lobe is more than 40dB down.

The Hamming window Characteristics are as follows:

Main lobe:

$$W_2(e^0) = 0.54L \qquad (3.37)$$

First null:

$$W_2\left(e^{j\frac{4\pi}{L}}\right) = 0 \qquad (3.38)$$

First sidelobe:

$$20\log_{10}\left(\frac{|W_2\left(e^{j\frac{5\pi}{L}}\right)|}{|W_2(e^0)|}\right) \approx -41\text{dB} \qquad (3.39)$$

Optimum window length depends on the necessity for model.
The main lobe width $(2 \times \omega_c)$ should not exceed F_0in order to resolve F_0 .
For example, for a Hamming window,

$$\frac{8\pi}{L} \leq \frac{2\pi F_0}{F_s} \qquad (3.40)$$

If you want to resolve the formants, main lobe should be wide enough to \smooth together" the pitch harmonics, but not so wide that it blurs the formant peaks.
If you want to look at rapidly changing events (e.g. stop release), L should be as short as possible (5-10ms max).

3.8 Causes of Speech Disorder (Articulatory Handicapped)

Point 3.2 explains how speech is produced and what the various vocal organs are responsible for generating understandable speech. If, for any reason, these organs suffer from any disease, speech will not be properly produced. In order to overcome the related problems, it is necessary to study the details of the causes of speech

disorder. If the speech is not properly produced that the person wants to make the sound of the words, it is said to be speech disorder or articulatory disorder. The following section explains causes of speech disorders. Many people with speech disabilities are conscious of what they want to say, but they are unable to express their feelings, which can lead to problems with self-esteem and depression. Articulatory disorder affects the different vocal organs such as vocal cords, muscles, nerves and other structures within the throat. Reasons may include vocal cord damage, brain damage, muscle weakness, respiratory weakness, strokes, polyps or nodules on the vocal cords and vocal cord paralysis. Speech disorder may develop due to certain medical or developmental conditions.

Following are some of the causes which may contribute to speech disorders are:

- Autism
- Genetic disorder
- Cancer related to throat
- Stroke
- Huntington's disease
- Dementia etc.
- Hereditary

Several indications may be present, depending on the cause of the speech disorder. Common symptoms experienced by human being are as follows:-

- Multiple noises that are most often seen in individuals who are stuttering
- Heaviness or talking with a raspy or serious tone
- Taking frequent breaks while delivering the speech
- Blinking several times during talking
- Find difficulties in communication due to visual frustration
- Presence of jerky head movements while speaking
- Elongation of words

3.9 Types of Speech Impairments

Articulation disorder, voice disorder and fluency disorder are the three basic types of speech impairments**.**

- Articulation disorder: An anatomical or physiological limitation in the skeletal, muscular, or neuromuscular support generates articulatory disorder which produces error in the speech production. These disorders consist of exclusion (ra for rat), replacement (ken for ten) and distortion (fffun for fun).
- Fluency disorder: Conversational disorders are difficulties with the rhythm and timing of speech which have been marked by hesitations, repetitions or prolongations of tones, syllables, words or phrases. Stuttering and cluttering are the two major types of fluency disorder. Stuttering consists of elongation, substitutions, repetitions, hesitations and whole verbal block. Cluttering produces due to speedily delivery of talk and shaky speech.
- Voice disorder: Voice disorders are problems caused by larynx disorders in the performance or use of one's voice. Voice disorders are characterized by irregular voice quality, volume, loudness, vibration and/or period development and/or absence.

Person’s ability is affected due to speech disability. Speech disorder is not same as language disorder. Voice disabilities prohibit people from making proper speech sounds, whereas communication problems impair a person's ability to remember vocabulary and comprehend what others speak to them. Nevertheless, both speech and language difficulties may make it more difficult for an individual to communicate his or her thoughts and feelings to others.

3.10 Statistical Representation of Speech Recognition System

A statistical representation of voice recognition system in simple equations that also include front end unit, model unit, language model unit, and search unit is shown in Figure: 3.6

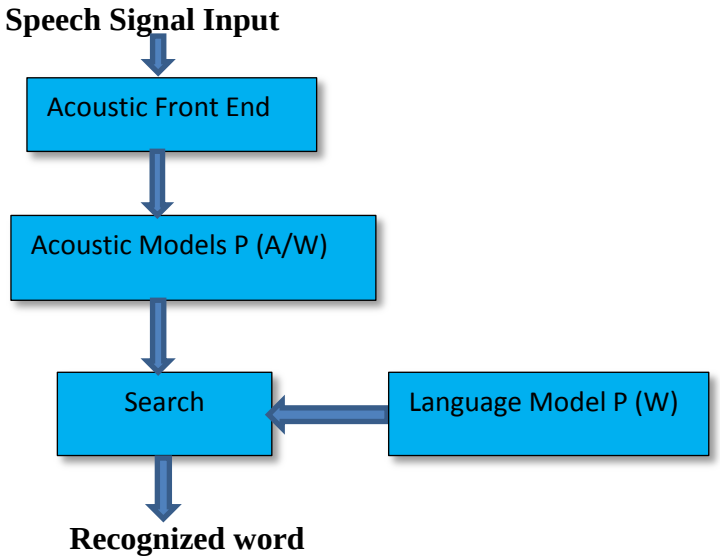


Figure 3.6: Statistical representation of voice recognition system

The standard approach to continuous speech recognition for large vocabulary is to assume a simple probabilistic model of speech output whereby a given word sequence, W, produces an acoustic observation sequence Y with probability P (W, Y).The aim is then to decode the word string based on the sequence of acoustic analysis, so that the decoded string has the highest likelihood of a posteriori (MAP).

3.11 Summary

The articulatory system is extraordinarily wonderful part of human speech production system that produces different sounds with different characteristics. Regardless of any circumstance the fault falls across this articulatory system, the tone of the voice is distorted and the human being suffers from speech disorder. Keeping in mind the objectives mentioned above, speech recognition and rectification algorithms have been developed and implemented in order to overcome the difficulties of articulatory speech by people with disabilities.

All problems relating to speech disorder and methods of speech recognition linked to speech disability have been discussed in Chapter 2. In previous studies, most researchers had good results for MFCC as a feature extraction technique. Various

classifiers are also used to identify the correct word. Previous studies on speech recognition for speech disorders show that there are different techniques proposed for speech recognition for articulatory disabled people. There are few limitations to each of these techniques. And therefore none of the techniques has yet been fully developed.

We intend to develop and implement an algorithm that provides better accuracy with less complexity of speech recognition. The research was conducted in two stages, namely,

- Speech Recognition without phoneme separation
- Speech Recognition and rectification with phoneme separation.

First data base is created for people with articulatory disorders in the first phase of the study. In this study, choice of the technique of extraction of features is very important. Different extraction techniques are used to achieve good accuracy of speech recognition and less word error rate. After implementation, the result shows that the MFCC features extraction technique provides better accuracy. The MFCC algorithm is implemented by different parameters and an appropriate set of parameters has been selected for further studies. After the choice of the feature extraction method, the accuracy of speech recognition is measured using various classifiers. Experimentation reveals that, among all the various classifiers, the k-NN classifier increases speech recognition performance by using 512 MFCC coefficients, 20 attributes and the hamming window.

In the second phase of the investigation, the separation of the phoneme by time domain segmentation is used to rectify speech recognition by keeping all parameters fixed as decided in the first stage. All these algorithms are implemented in next chapter.

Chapter 4

DEVELOPMENT AND IMPLEMENTATION OF SPEECH RECOGNITION AND RECTIFICATION FOR ARTICULATORY HANDICAPPED PEOPLE

4.1 Overview

The voice recognition system consists of extraction features, acoustic model, lexicon model and language model. In order to perform speech recognition, each of these modules could be designed separately and merged together. From a statistical point of view, ASR is the task of ending the most likely sequence of words W corresponding to our O observation. Following the formula that uses the Bayes rule,

$$P(W|O) = \frac{p(O|W)P(W)}{p(O)} \quad (4.1)$$

The Bayes rule breaks the model into two parts: the acoustic model p (O) and the language model P (W). P (W) is the word sequence, W probability and p (O) is the later probability density function value. ASR's becomes,

$$\arg \max_{w} P(W|O) = \arg \max_{w} p(O|W)P(W) \quad (4.2)$$

Assuming that P (O) is true, add a Q phone string to further decompose the first element as hidden variables:

$$\arg \max_{w} P(W|O) = \arg \max_{w} \sum_{Q} p(O|Q)\, P(Q|W)P(W) \quad (4.3)$$

And these three components correspond respectively to the modeling of acoustics, lexicon and language [58].

4.1.1 Lexicon

A lexicon, also known as a dictionary, is a map of their pronunciations in written word phones. This portion tends to help narrow the gap between embedded words and their acoustics. Using a lexicon, it allows words that do not appear in a training set to be recognized. Phone is a speech segment that has distinct physical or perceptual properties and serves as the basic unit for phonetic speech analysis. The number of

phones can range from 40 to a few hundred for a typical spoken language. In general, phones are either vowels or consonants and can last from 5 to 15 frames (50 to150ms). During analysis, the lexicon is used to restrict the search path for validating phone strings and translating phone strings into words [58]. Figure 4.1 shows models for Speech Segment.

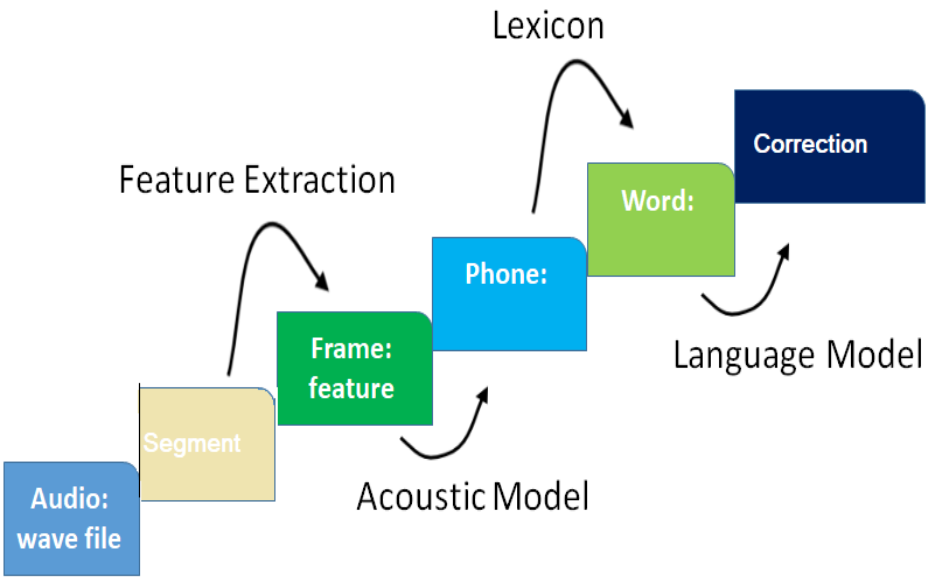


Figure 4.1: Models for Speech Segment (source: Yanzhang He, 2015)

Table 4.1 presents the lexicon table of digits zero to 10, which provides information on phonemes.

Table 4.1: Lexicon Table

Digits	**Phonemes**
One	w ax n
Two	t uw
Three	th r iy
Four	f ow r
Five	f ay v
Six	s ih k s
Seven	s eh v eh n
Eight	ey t
Nine	n ay n
Ten	t eh n
Zero	z iy r ow

Language Model (LM) is a probability function over word series, P (W) is decomposed as,

$$P(W) = P(\omega_1, \omega_2, \dots, \omega_n) = P(\omega_1)P(\omega_2|\omega_1) \dots P(\omega_n|\omega_1, \dots, \omega_{n-1}) \quad (4.4)$$

And each of these probabilities could be estimated using maximum probability methods by checking the count of word sequences $c(\omega_1, \omega_2, \ldots, \omega_n)$ in the training corpus.

$$(\omega_n|\omega_1, \ldots, \omega_{n-1}) = \frac{c(\omega_1,\omega_2,\ldots,\omega_n)}{c(\omega_1,\omega_2,\ldots,\omega_{n-1})} \quad (4.5)$$

Owing to all the sparsity of training data, traditional ASR language models make the Markov implication of k-order. i.e.

$$(\omega_n|\omega_1, \ldots, \omega_{n-1}) = P(\omega_n|\omega_{n-k}, \ldots, \omega_{n-1}) \quad (4.6)$$

For actual ASR implementations, k = 2 and k = 3 are the most common configurations, and corresponding LMs are referred to as "bigram template" and "trigram design" respectively

.4.1.2 Acoustic Modeling

Acoustic modeling is a critical component of speech recognition systems. Simply put, acoustic modeling is the task of creating statistical models or deep learning models to estimate the probability of output p(X).

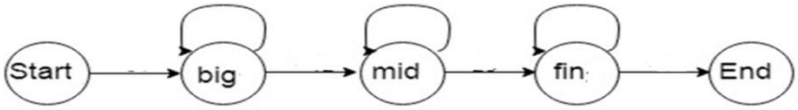
In a generational approach, the typical distribution of X and Q, p(X) is modeled. In a biased approach, the resulting distribution p (Q) is specifically modeled. In a hybrid approach, both a generative model and a discriminative model are used.

4.1.3 Phoneme

As people speak, air passes from the lungs and creates sound through the oral cavity and nasal cavity. Usually called phonemes, vowels and consonants are formed by these sounds. The phonemes are combined to form sentences in order to form words and words. In a particular language, a phoneme distinguishes one word from another.

For example, the thumb and dumb sound patterns in most English dialects are two separate words that are separated by replacing one phoneme, /everything/, with another phoneme, /d/. Two words like this, differing in context by comparing the form of a single phoneme called a minimal pair. A total of 18 phonemes are present in 0 to 10 digits [59]. The model of Speech Recognition using Phoneme Level, Word Level shown in figure 4.2

Phoneme Level



Word Level

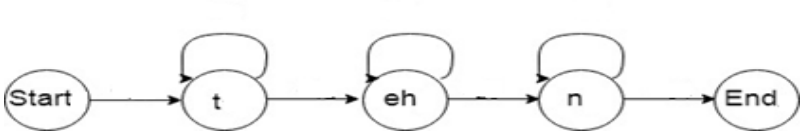


Figure 4.2: Model of Speech Recognition using Phoneme Level, Word Level (source: cvimala)

4.2 Structure of Speech Recognition System

All algorithms have been developed using MATLAB 2017b and are tested on database generated for articulatory handicapped people. The framework is carried out in two broad categories, namely,

- Speech Recognition without phoneme separation
- Speech Recognition and rectification with phoneme separation.

Figure 4.3 shows Structure of Speech Recognition systems for articulatory handicapped people without phoneme separation.

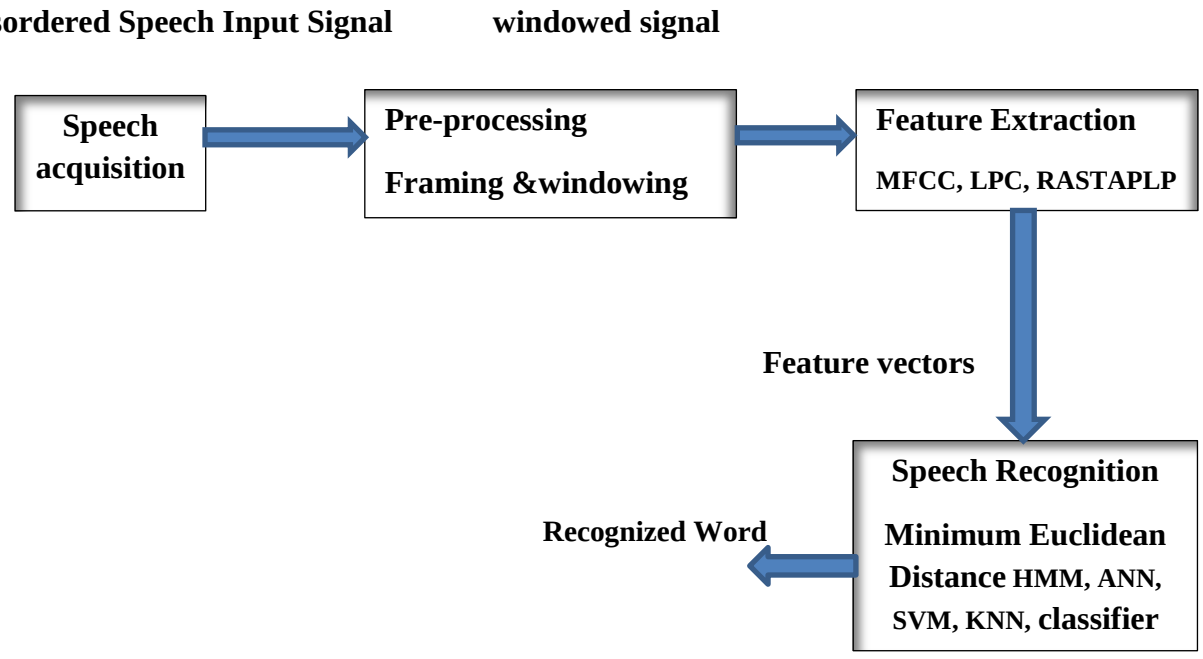


Figure 4.3: Structure of Speech Recognition systems for articulatory handicapped people without phoneme separation

As shown in Figure 4.3, speech recognition was carried out using four significant steps, namely speech acquisition, pre-processing, extraction of features and speech recognition.

4.2.1 Speech Acquisition

Any speech recognizer's first stage is to acquire voice data. It is important to prepare the speech corpus for both training and testing of the process. A microphone detects the acoustic waves produced by the vocal tract system during speech acquisition. The articulation generates sound waves in the human speech production system, and Ear transmits it to the brain for processing. As people speak, air passes through the oral cavity and nasal cavity from the lungs and creates noise. Such sounds involve voices or consonants, usually called phonemes. Phonemes are merged to shape words, and phrases are joined together to create sentences.

The three key components to be carefully considered during the speech acquisition process were:

- Sample Rate - How often voltage values are recorded,
- Bits per sample - How accurate it is to record the value, and
- Number of channels - mono or stereo.

Human ears can hear frequencies from 20 Hz to 20 kHz and human voices range from 300 Hz to 3,400 Hz. A minimum bandwidth of 4 kHz is required to convey information to each other.

In this proposed system, the criteria used for speech acquisition are summarized under Table 4.2.

Table 4.2 Specification used for Speech Acquisition

Specifications	**Metaphors**
Format of Input File	Wav
Sampling Rate	44,100 Hz
Format of Sampling	16 bits
Input Channel	Mono
Digitization technique	Pulse-Code Modulation(PCM)
Software for Recording	NUENDO 4

4.2.2 Speech Pre-processing

Speech Pre-processing is seen as an important step towards the development of a robust speech recognition system. This increases the reliability and performance of speech recognition system. Signals are usually pre-processed before further analysis. It requires significant measures such as digitization, sorting, pre-emphasis scanning, framing, and windowing. Such preprocessing methods are explained in the following sections.

4.2.3 Digitization and Sampling

Converting an analog signal into a digital signal is the first step of speech pre-processing. Digitized speech as a string of numbers represents a voice signal. A digitized signal is sampled at a particular time by calculating its amplitude. The rate of sampling is an important factor because it calculates the number of samples per second. In each phase, a minimum of two samples is needed to accurately measure the waveform. Another sample is used to measure positive signal parts and another sample is used to measure negative signal parts. 44.1 kHz sampling frequency has been used in this research work.

4.2.4 Pre-emphasis Filtering

In this step, the input signal is applied to the pre-emphasis filter. It is used to better distinguish low-energy, high-frequency consonants from high-energy and low-frequency voices. High-frequency energy enhancement provides the acoustic model with more data and increases the efficiency of recognition. The resulting signal closely matches a higher frequency boost to the original signal.

4.2.5 Framing

The ability of a human ear is not to respond to a very rapid change in speech content. It is therefore necessary to divide the speech data into short frames with a certain number of samples prior to analysis. Although the speech signal is a constantly changing time variable component, it becomes statistically stationary in a short period of time. Selecting the correct frame size is a very important step in speech signal processing due to the trade-off between time and frequency resolution. The smaller frame size will not provide an accurate spectral estimation, and therefore the frequency resolution will also decrease. If the frame size is larger, the frequency

variable can shift too much throughout the frame. Therefore, the reliable spectral properties of the signal cannot be extracted. By taking into account the above factor, frame sizes between 10-30 ms are normally chosen for speech pre-processing applications.

4.2.6 Windowing and Overlapping of frames

Framing incorporates distorted spectral information in segmented voice, so that all framing samples are combined by a window weighing function to reduce this distortion. Gradually, it can stop any changes between frames. Framing incorporates distorted spectral information in segmented voice, so that all framing samples are combined by a window weighing function to reduce this distortion. Gradually, it can stop any changes between frames (Priyanka Prakash Katule, and Pathak, B.V, 2013). In general, the window size should be longer than the frame duration to avoid any loss of information during transitions between the frames. To prevent missing data, repetition of frames is used to maintain continuity between the previous frames. Figure 4.4 displays the overlapping frame model (Bibek Kumar Padhy, 2009).

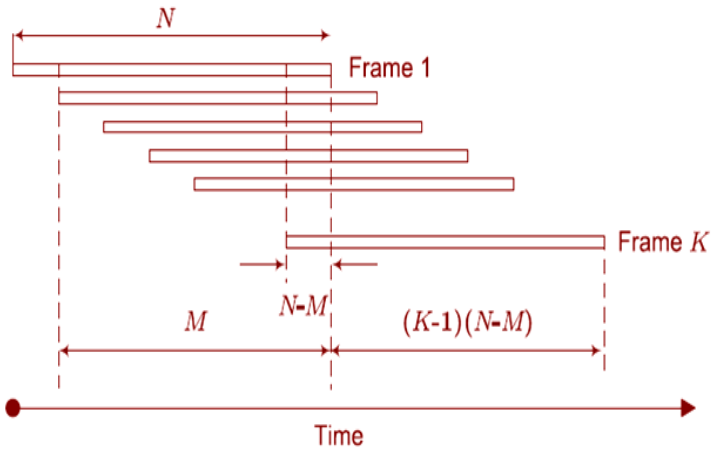


Figure 4.4: Overlap structure of frames (source Bibek Kumar Padhy, 2009)

N is the number of samples per frame. M samples overlapped with the preceding frame and K is the number of frames used to average. N-M filtered samples will be obtained after processing the new frame. Such overlapping regions can use frame size ranges ranging from 0 to 75 percent. For studies, 60 percent overlap and 25ms window size is used in this research work. This is because, over 20-25msec time window, human speech can be considered relatively stationary. In signal processing, different types of windows are available. Four types of window functions, namely

rectangular, triangular, hanning and hamming, were used in this proposed system. A good window feature range has a large main lobe in its transition and low side lobe rates. The hamming window from above all sort of window is the best choice.

4.3 Feature extraction techniques

The purpose of the extraction function technique is to transform the raw signal into characteristic vectors that emphasize the properties of expression. It also improves the system's performance. The steps involved in feature extraction shown in Figure 4.5 which also indicates mechanism for overlapping of frames (Gonzalez J, Lopez-Moreno,2010)

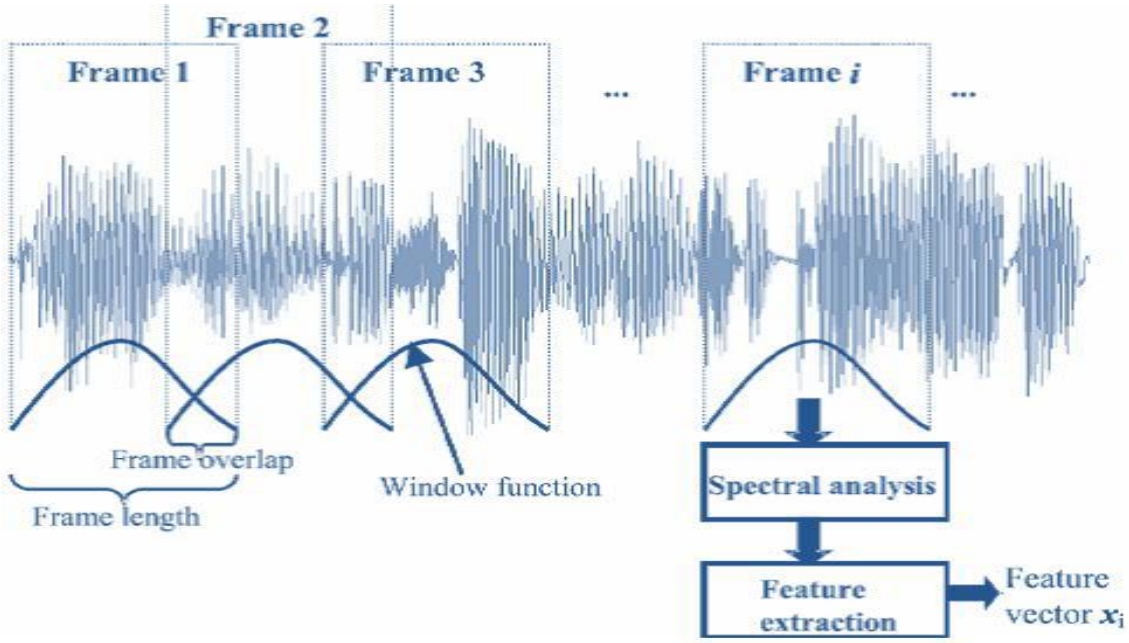


Figure 4.5: Process of Feature Extraction in Overlapping Frames (source-Gonzalez J, Lopez-Moreno)

In essence, speech signal consists of two types of attributes, such as time domain features (temporal features) and frequency domain features (spectral features).commonly used feature extraction techniques. Temporal features are the energy of signal, zero crossing rate, maximum amplitude, minimum energy, etc. and spectral features are fundamental frequency, frequency components, spectral centroid, spectral flux, spectral density, spectral roll-off, etc. Notes, pitch, rhythm and melody

can be determined by using these features. The most common attributes used in speech recognition system are as follows.

- Mel Frequency Cepstral Coefficients (MFCC)
- Linear Predictive Coding (LPC)
- Relative spectral Perceptual Linear prediction (RASTA PLP)

4.3.1 Mel Frequency Cepstral Coefficient (MFCC)

The argument about speech is that a human-generated sound is produced by the vocal tract shape along with tongue, teeth, etc. This form determines the resulting sound. If this shape is correctly calculated, the phoneme being produced should be accurately represented. The vocal tract shape expresses itself in the short-term power spectrum envelope, and MFCCs 'job is to accurately represent this envelope.

The following Figure 4.6 describes the different MFCC blocks.

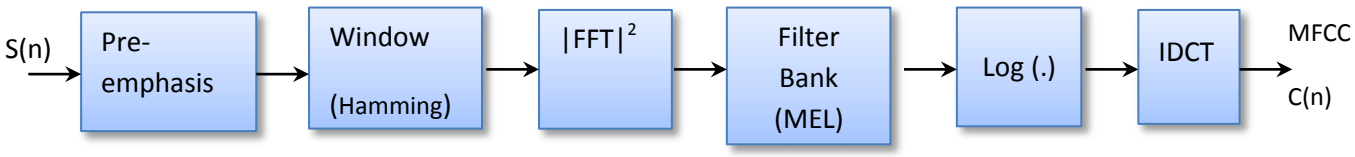


Figure 4.6: Block diagram of MFCC

Algorithm to find MFCC coefficient

Step 1: Pre-process the voice output signal that is speech signal

Step 2: Split the speech signal into short frames

Step 3: Window the short frame consists of a speech signal

Step 3: Calculate the periodogram estimation of the power spectrum for each frame

Step 4: Apply the Mel filter bank to the power spectrum, add each filter's energy.

Step 5: Take the logarithm of all filter bank energies.

Step 6: Determine the log filter bank's DCT

Step 7: Choose a DCT coefficient to achieve the maximum accuracy of the digit.

4.3.1.1 Pre-Emphasis

The first step is to add a pre-emphasis filter to the signal to amplify the frequencies. A pre-emphasis filter is effective in several ways:-

- Adjust the frequency spectrum, since high frequencies tend to be smaller than lower frequencies;
- Prevent numerical problems during the process of transformation in Fourier
- Enhances the SNR (Signal-to-Noise Ratio).

The pre-emphasis goal is to account for the heavy-frequency component lost during the development of human sound. The higher frequency energy of the signal will be highlighted in this process.

Pre-emphasis uses 1st order FIR high pass filter, whose transfer function is,

$$H[z] = 1-a*Z^{-1} \qquad (4.7)$$

Where, $0.95 \leq a \leq 1$

After pre-emphasis, the signal has the following form in the frequency domain shown in Figure 4.7 (a), (b) and (c)

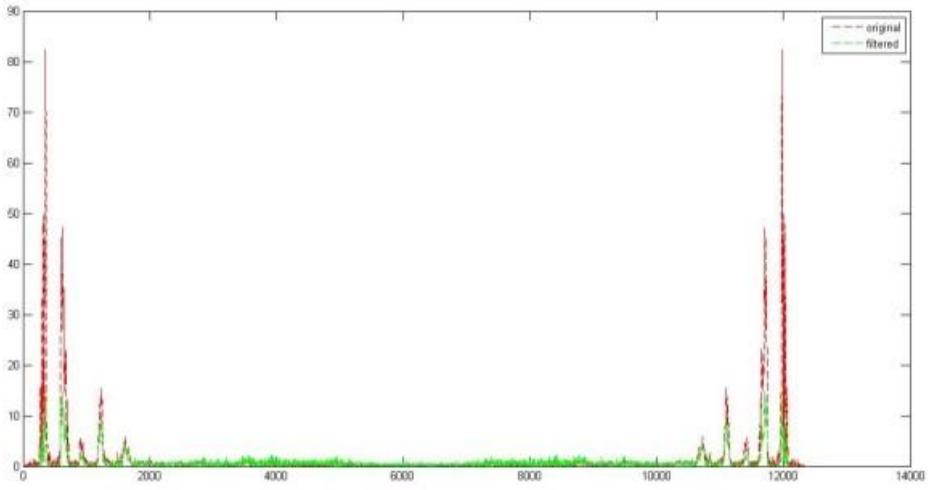


Figure 4.7(a): The signal after pre-emphasis has the above form in the frequency domain below cutoff frequency of FIR high pass filter

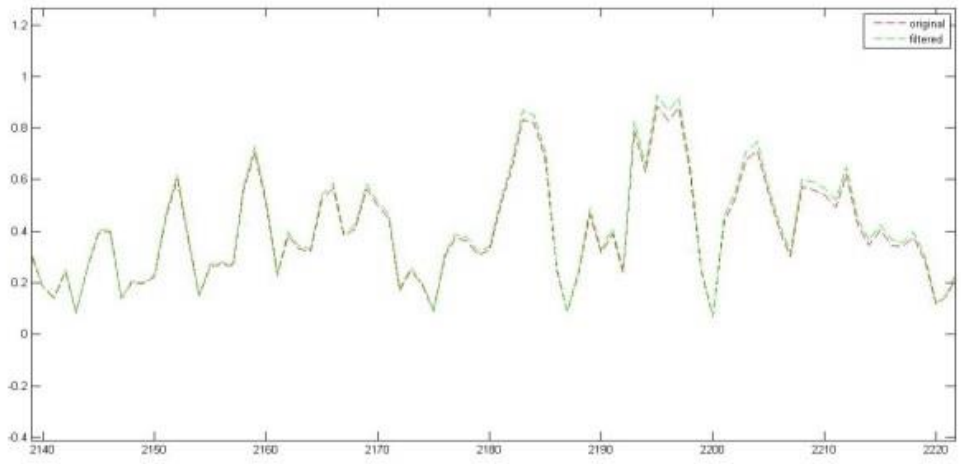


Figure 4.7(b): The signal after pre-emphasis has the above form in the frequency domain above cutoff frequency of FIR high pass filter

Above 2 KHz frequency (cutoff of 1st order FIR high pass filter, the speech signal emphasized that is boosted which is suppressed during production of sound process.

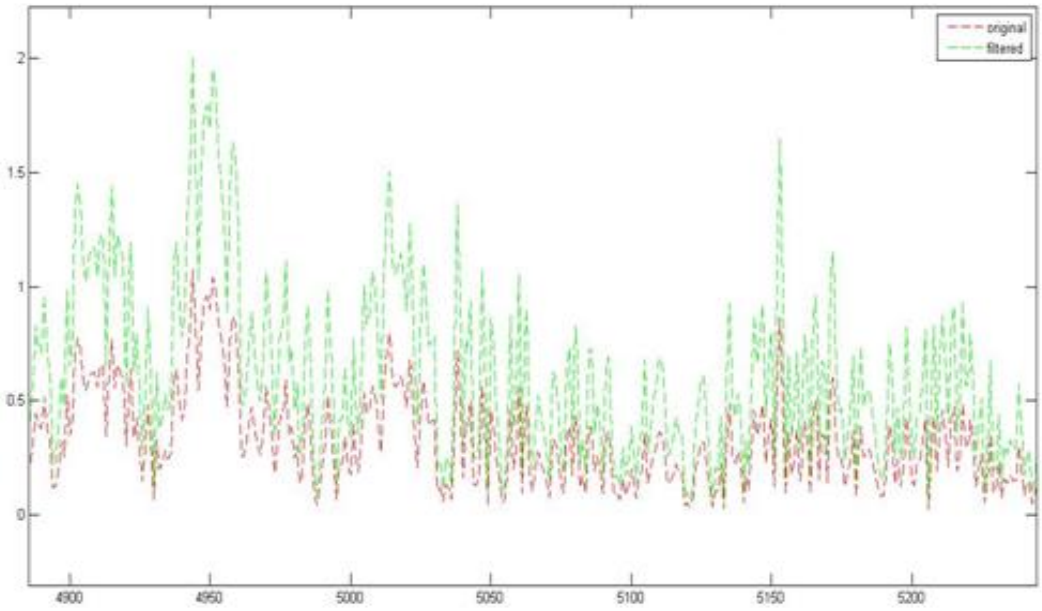


Figure 4.7(c): Emphasized High frequency speech signal after pre-emphasis

Figure 4.7(b) & (c) shows that amplitude of filtered signal (shown in green color) is greater than that of the amplitude of original signal (shown in red color).

- **Selection of Filter Coefficient 'a' for Pre-emphasis Filter**
  - When the "trace" unit circle point is close to zero, the corresponding frequency will be attenuated. Alternatively, if a point on the unit circle is near a pole, it will be amplified.

- Trace the circle of the unit by setting z= e^{jw}.
- There will be a "zero" if a = z, setting ' a ' to 0.9 or 0.97 respectively sets the "zero" to 0.9 or 0.97. This filter can attenuate frequencies close to ω=0.
- The value a=0.97 attenuate lower frequencies by more than a=0.9.
- The option of relevance ' a ' depends on the nature of the device or streams used to store and relay the message signal.

- **Magnitude Plot of Pre-emphasis Filter**
  - For the frequency response, substitute $exp\left(-j\frac{2\pi f}{f_s}\right)$for z. This yields,

$$H(f) = 1 - a\ exp\left(-j\frac{2\pi f}{f_s}\right) \qquad (4.8)$$

  - Magnitude response is taking modulus us of equation 4.8
  - By plotting this function and we will see that it is a high pass response with a maximum at 1/2 of sampling frequency'

$$|H(f)| = 1 + a^2 - 2a\cos\left(\frac{2\pi f}{f_s}\right) \qquad (4.9)$$

The filter coefficient 'a' affects the gain, roll-off and DC component of speech signal. An Effect of Pre-emphasis coefficient factor "a" on speech recognition accuracy shown in figure 4.8

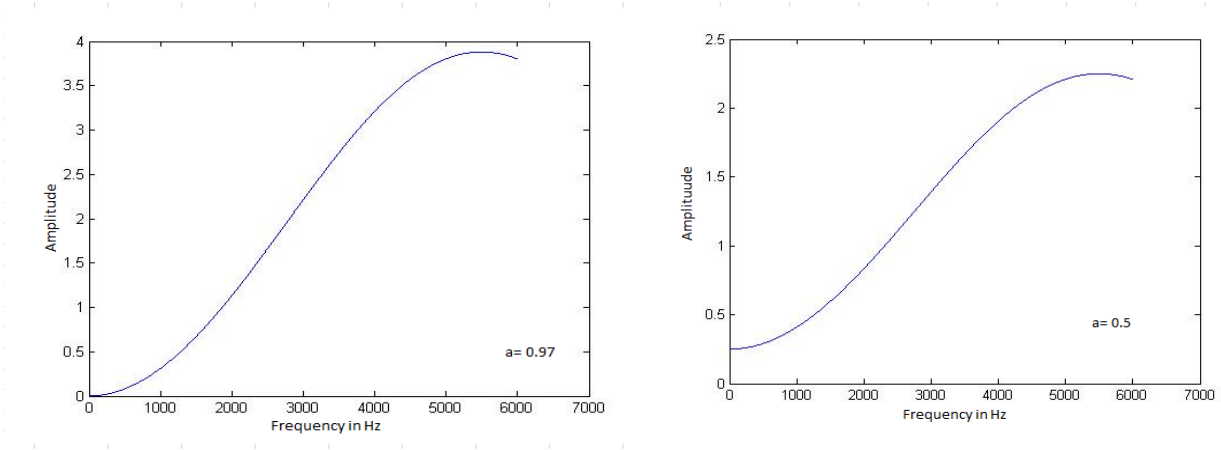


Figure 4.8: Effect of Pre-emphasis coefficient factor "a" on speech recognition accuracy

4.3.1.2 Frame and Selection of Frame length

Divide the signal into short-term frames after the pre-emphasis. The rationale for this step is that frequencies shift in a signal over time, but transmitting the Fourier over the entire signal in most cases does not make sense as it removes the signal's frequency contours over time. But in a very short time, frequencies are stationary in a signal. As the human speech signal is non-stationary, this signal state can be eliminated by means of short-term analysis. Short-term spectral analysis is carried out using Fourier transform short-term analysis

$$S_n(e^{j\omega}) = \sum_{m=0}^{L-1} s(m)w(n-m)e^{-j\omega m} \quad (4.10)$$

Where, $S_n(e^{j\omega})$ is Fourier transform of the windowed signal $s(m)w(n-m)$, n reflects an evaluation of the window change in the number of samples.

The frequency resolution of the spectral analysis Δf is associated with the length of the analyzed signal N and the frequency of sampling Fs. Taking into account the length of the analyzed signal, the length of the analysis window tw is determined.

$$\Delta f = \frac{Fs}{N} = \frac{Fs}{tw * Fs} = \frac{1}{tw} \quad (4.11)$$

A longer analysis window uses a higher frequency resolution and a lower time resolution.

Therefore, the 20 ms analyzing window gives us 50 Hz frequency resolution. Shorting the window analysis to 10 ms gives a higher time resolution, but the frequency resolution will be reduced to 100 Hz, which can stress the low frequency pitch-based speech analysis process and its harmonics. The spectral analysis outcome is therefore ultimately influenced by the width of the frame [60]. If the analysis window is too broad, the study cannot accurately represent differences in speech signal characteristics. This implies that for all cases there is no clear and valid window length quality, as the pitch time differs within speakers. It's higher for women or kids, lower for males. The pitch time ranges between 3.3 and 16.6 ms (pitch frequency varies between 60 and 300 Hz). The analyzed voice segment may become non-stationary by extending the length of the analysis window. If the length of the analysis

frame is too shortened, some signal characteristics will be lost. In addition, if we shorten the analysis window until it is shorter than the pitch period (2–3 ms) we may miss registering the pitch peaks [70]. In this research, different frame length analysis windows were used to estimate the spectrum. The size of the frame may depend on the speaking rate, including specific sounds [70].

4.3.1.3 Selection of frame overlapping

Another important aspect of research is the extent to which the window (frame) change overlaps or analyzes. The size of the window change determines the specificity of speech dynamic data. The lower frame shift we use, the more information we can get from the dynamics of voice. Such testing can take longer, however, and does not necessarily mean a higher rate of speech recognition. The variance is usually chosen equivalent to half or one third of the duration of the study period [70].

The width of the speech signal and the evaluation range, the frame shift, is linked to:

$$M = \frac{N - L}{\Delta L} + 1 \qquad (4.12)$$

Where M is the amount of frames obtained in the speech signal, N is the width of the signal, L is the duration of the observation period, ΔL is the shift of the analysis window. Therefore, the increase or decrease in the length of the analytical window will not necessarily affect the number of extracted frames to be analyzed. Practically consideration is given to L<<N. Therefore, the main criterion for determining the size of the study window is the resolution of the frequency.

The change of the frame specifies the number of frames obtained. The smaller the window change, the longer it takes to evaluate the signal, the more signal frames are extracted. The size of the frame shift ΔL is directly related to the number of M analytical frames, e.g. by reducing the frame shift twice, the number of frames will be twice lower, so analytical time should also be decreased twice. This is important on devices with limited calculation resources for speech recognition. Frame shift control can save time and power in the calculation. In this study, overlap is set for the analysis window for different sizes (50 percent, 60 percent, and 70 percent).

4.3.1.4 Selection of window type

Most digital signals are infinite or large enough to be able to manipulate the dataset as a whole. It is also difficult to statistically analyze sufficiently large signals, as statistical calculations require that all points be available for analysis. To avoid these problems, it is necessary to analyze virtually small subsets of the total data through a process called windowing. The speech signal is stationary for a short period of time, so that the Fourier transform technique is used for speech analysis. This results in spectral leakage due to the abrupt truncation of the Fourier series. Frame windows are used to reduce the spectral effect and to smooth the signals for FFT computing. Hamming, Hanning, rectangular and triangular are commonly used window functions. Depending on the following characteristics, the type of window is selected. Following Table 4.3 displays the features of different windows.

Table 4.3 Different Window Features

Window Type	Peak Sidelobe Amplitude (Relative, dB)	Approximate Width of Main Lobe	Peak Approximation Error, $20\log(\delta)$ (dB)
Rectangular	-13	$\frac{4\pi}{M+1}$	-21
Bartlett	-25	$\frac{8\pi}{M}$	-25
Hann	-31	$\frac{8\pi}{M}$	-44
Hamming	-41	$\frac{8\pi}{M}$	-53
Blackman	-57	$\frac{12\pi}{M}$	-74

Where, M is window length. The table above shows that the width of the main lobe is a function of window length, but the stop band attenuation depends on the type of window. Thus, when selecting the type of window, a trade-off between the width of the transition band and the stop band attenuation must always be considered. This work carried out experiments by taking into account all types of windows. This shows that the hamming window yields the most accurate results.

- **Hamming Window**

A Hamming window multiplies the resulting frame to mitigate the spectral leakage effect. The Hamming window has almost zero values towards both ends, ensuring signal continuity in successive frames. A Hamming window has the following form:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right), 0 \leq n \leq N-1 \qquad (4.13)$$

Where, M is number of samples per window. Behavior of hamming window is shown in figure 4.9.

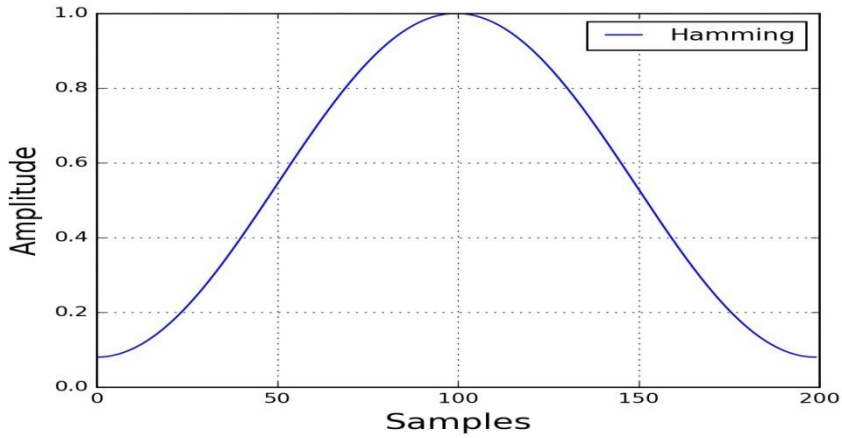


Figure 4.9: Nature of hamming window

The window function is chosen based on three performance parameters such as,

- **Sidelobe cancellation**

The Hamming window's maximum side lobe attenuation is −43 dB, which is more than any other form of window except Blackman window

- **Worst Case Processing Loss (WCPL)**

A system's worst case processing loss is the sum of the processing loss and the system's scalloping loss. A system's worst case performance failure is the amount of processing loss and the system's scalloping loss. The scalloping failure of an N-length frame s (k)

$$SL = \left| \frac{\sum_{k=o}^{M-1} s(k) e^{-\frac{j\pi k}{M}}}{\sum_{k=o}^{M-1} s(k)} \right| \qquad (4.14)$$

The scalloping loss is the coherent gain of a frequency half a bin from a component on the discrete Fourier transform divided by the coherent gain of the window.

- **Equivalent Noise Bandwidth (ENB)**

Bandwidth of Equivalent Noise (ENBW) compares a window to an ideal, rectangular time-window. It is the bandwidth of the frequency-domain shape of the rectangular window that passes the same amount of white noise energy that the other window defines as the frequency-domain shape.

4.3.1.5 Fourier-Transform and Power Spectrum

The FFT was structured to convert the convolution of the glottal pulse and the response of the vocal tract to the time domain H[n].If X (w), H (w) and Y (w) are respectively X (t), H (t) and Y (t) Fourier transforms. Spectral analysis reveals that various timbres in the speech signal lead to different energy distribution over frequencies. Therefore, the FFT is performed to obtain the frequency response of each frame [38].

Where M is number of samples and then uses the following equation to calculate the power spectrum (periodogram):

$$P = \frac{|FFT(x_j)|^2}{M} \qquad (4.15)$$

Where, x_j is the j^{th} frame of signal x.

4.3.1.6 Filter Banks

The next stage in the computation of filter banks is to add triangular filters, 40 filters that are used to procure the frequency bands on the Mel-scale to the power spectrum. The goal of the Mel-scale is to mimic the interpretation of sound by non-linear human ears by being more discriminating at lower frequencies and less discriminatory at

higher frequencies. The following formula shows the correlation between Hz and Mel-Scale output signal frequency:

$$m(f) = 2595log10\left(1 + \frac{f}{700}\right) \quad (4.16)$$

$$f = 700(10^{\frac{m}{2595}} - 1) \quad (4.17)$$

Every filter in the filter bank is triangular with a center frequency response of 1 and decreases linearly to 0 until it approaches the center frequency of the two neighboring filters with a response of 0, as shown in figure 4.10.

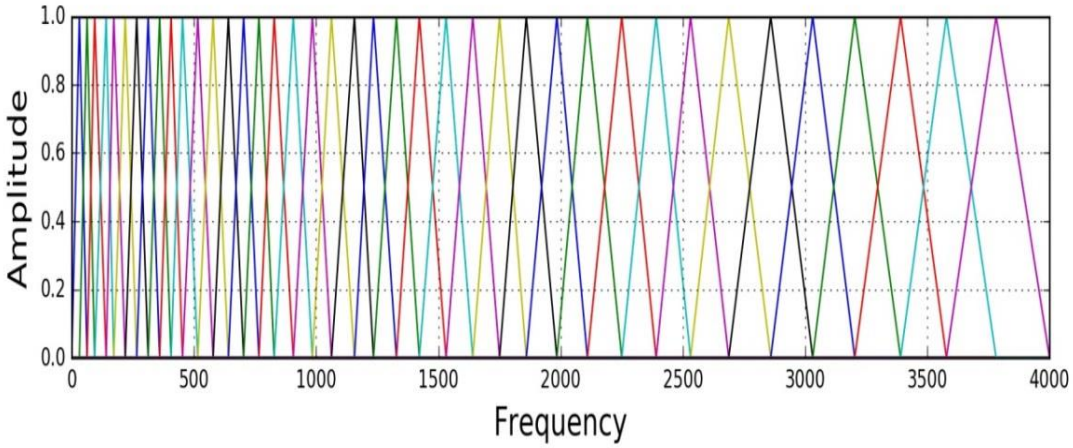


Figure 4.10: Filter bank on a Mel-Scale (source:2016, Haytham Hayek)

This can be modeled by the following equation (source: 2016, Haytham Hayek)

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k < f(m) \\ 1 & k = f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \le f(m+1) \\ 0 & k > f(m-1) \end{cases} \quad (4.18)$$

The spectrogram shown in figure 4.11 is produced after adding the filter bank to the energy spectrum of the signal:

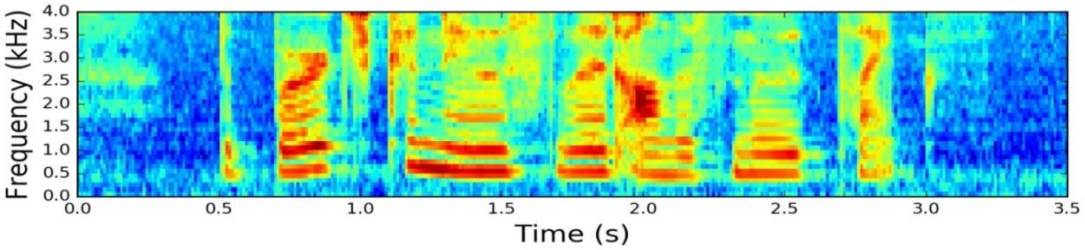


Figure 4.11: Spectrogram of signal (source:2016, Haytham Hayek)

- **Discrete Cosine Transform**

The step of MFCC is to calculate the DCT energy log filterbank. There are two main reasons why this has been done. Because all of the filter banks overlap with the energies, as the filter banks are closely correlated. The DCT decorrelates the energies that can be used through diagonal covariance matrices to model the features in e.g. a KNN classifier. But note that it will retain only 20 of the DCT's 40 coefficients. This is because the higher DCT coefficients reflect rapid changes in the energies of the filterbank and it turns out that these rapid changes actually reduce ASR efficiency, thereby improving performance by lowering them.

4.3.1.7 Observations

By changing the order of the FIR high pass filter the effect of pre-emphasis has been observed, it is found that the results do not change. Changes in frame size have been observed to affect speech recognition accuracy. It has been found that the percentage of overlapping of frames also affects accuracy. Window type also affects the performance of the MFCC feature extraction method and it has been observed that widow selection is a function of parameters such as sidelobe cancellation, Worst Case Processing Loss and Equivalent Noise Bandwidth.

4.3.1.8 Linear Predictive Coding (LPC)

The method used to extract the attribute is Linear Predictive Coding (LPC). The linear method of prediction provides a reliable, fast and accurate method for estimating parameters characterizing the linear time-varying process representing vocal tract. Higher LP filters should be used to model the prediction of a signal, whereas the lower order LP filter is useful to obtain formant frequencies. LP filter order can be taken from 8 to 14.An all-pole LPC model provides a good approximation to the spectral vocal tract envelope in a voiced frame rather than an unvoiced frame (Karthikeyan Natarajan et al., 2008). A given speaking sample at time n, s(n) can be approximated in the LPC model as a linear combination of past p speaking samples as given in Equation (4.19).

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + - - + a_p (s-p) \quad (4.19)$$

It is assumed that the coefficients a_1, a_2, ... a_P over the frame are constant. LPC's main goal is to reduce the amount of square differences between an original voice signal and an approximate voice signal to provide a unique set of predictor coefficients. To derive these LPC coefficients, the speech signal is initially disrupted in frames of n samples and each frame is compounded by a hamming window. Subsequently, short-term self-correlation analyzes were conducted to define a significant frame energy factor for speech detection (Bibek Kumar Padhy, 2009). The Levinson Durbin Recursion algorithm is then implemented to derive the coefficients of the predictor and then converted into Q Cepstral coefficients, which are weighted by an enlarged sine window (Deller et al., 2000). Ultimately, the vectors of measurement are derived. Ultimately, the vectors of measurement are derived. In this research, 12 Cepstral coefficients were extracted from the feature vector.

The steps involved in the extraction of the LPC feature shown figure 4.12.

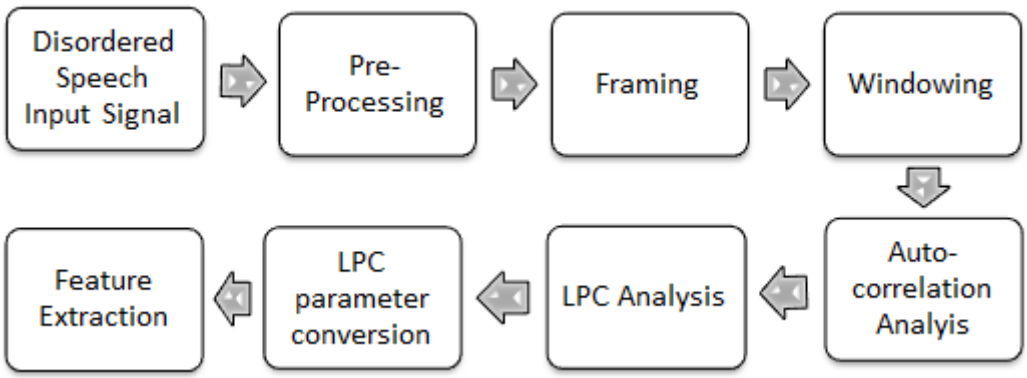


Figure 4.12: Steps for the extraction of LPC features

The LPC algorithm is outlined as follows:

1. Pre-emphasis

The speech signal $s(n)$ is put through a low-order digital system to flatten the signal spectrally and then render it less sensitive to finite precision effects in signal processing. The equation 4.20 provides relationship between the pre-emphasize network output and input s(n) to the network by the equation difference:

$$\tilde{s}(n) = s(n) - \tilde{a}s(n-1) \qquad (4.20)$$

b) Frame Blocking

The output of pre-emphasis is $\tilde{s}(n)$, blocked into frames of N samples, with adjacent frames being separated by M samples. If $x_l(n)$is the l^{th} frame of speech, and there are L frames within entire speech signal, then

$$x_l = \tilde{s}(Ml + n) \qquad (4.21)$$

Where, n = 0, 1… N – 1 and l = 0, 1…L – 1

1. Windowing

The signal discontinuities at the start and end are reduced due to windowing. If the window is defined as,$w(n)$ 0 ≤ n ≤ N– 1, the result of windowing is:

$$\widetilde{x_l}(n) = x_l(n)w(n) \qquad (4.22)$$

Where, $w(n)$ is between 0 ≤ n ≤ N – 1

Typical window is the shape of the Hamming window

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \qquad (4.23)$$

Where, $w(n)$ is having range 0 ≤ n ≤ N – 1

c) **Autocorrelation Analysis**

The next stage is to auto-correlate every frame of a windowed signal to provide,

$$r_l(m) = \sum_{n=0}^{N-1-m} \widetilde{x_l}(n)\widetilde{x_l}(n+m) \qquad (4.24)$$

Where, m = 0, 1. . .p And p is the order of LPC analysis

- **LPC Analysis**

Growing frame uses Durbin's method to translate p+1 autocorrelations into LPC parameters. This can be done by using the following equation:

$$E^{(0)} = r(0) \qquad (4.25)$$

$$k_i = \frac{r(i) - \sum_{j=1}^{i=1} \alpha_j^{i-1} r(|i-j|)}{E^{i-1}} \; 1 \le i \le p \qquad (4.26)$$

$$\alpha_i^{(i)} = k_i \quad (4.27)$$

$$\alpha_i^{(i)} = \alpha_j^{i-1} - k_i \alpha_{i-j}^{(i-1)} \; 1 \le j \le i-1 \quad (4.28)$$

$$E^i = (1 - k_i^2) E^{i-1} \quad (4.29)$$

Solving equations 4.29 recursively for i = 1, 2,…, p, the LPC coefficient, a_m, is given as

$$a_m = \alpha_m^{(p)} \quad (4.30)$$

- **Conversion of LPC Parameter into Cepstral Coefficients**

The following formulas are used to derive LPC parameters from the set of LPC coefficients,

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) . c_k . a_{m-k} \qquad 1 \le m \le p \quad (4.31)$$

$$c_m = \sum_{k=m-p}^{m-1} \left(\frac{k}{m}\right) . c_k . a_{m-k} \qquad m > p \quad (4.32)$$

In this work, by varying the order P of LPC coefficients, speech recognition accuracy is calculated.

4.3.1.9 Observations

The order of LPC analysis was found to influence the word error rate of speech recognition system. This study also found that the MFCC technique is more robust than the LPC feature extraction method.

4.3.1.10 Relative spectral Perceptual Linear Prediction (RASTA PLP)

The term RASTA is derived from the terms Relative Spectra. In the estimation of the critical band spectrum, the RASTA technique applies a bandpass filter to each spectral component. This filtering illustrates spectral differences between frame-to-frame frequencies between 1 and 10 Hz. Before applying the bandpass filter, log RASTA takes the usual logarithm of each. If we hear a sound, our human ear perceives it. In this procedure, this sensory feature of the human ear is captured. The

energy range of the speech signal is transformed to a bark scale similar to the sensory model of the human ear.

In the figure.4.13 the steps involved in calculation of features using RASTA-PLP is shown.

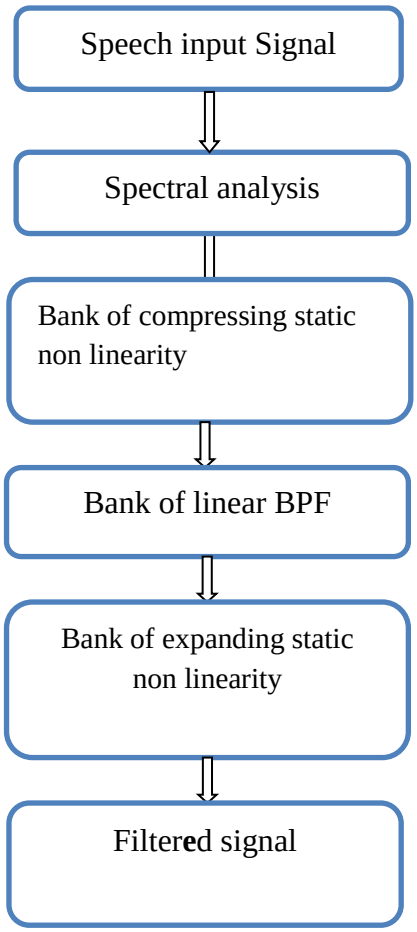


Figure 4.13: Steps involved in extracting features of speech signal using RASTA PLP feature extraction technique.

RASTA is a different technique which applies the band-pass filter to the energy of each frequency sub band in order to smooth over short-term noise fluctuations and to eliminate any continuous offset arising from stationary spectral coloring in the speech stream. The Rasta function includes the energy pitch, gain and length and the coefficients of short-term noise. The Rasta filter is used by the log spectral or Cepstral domain. In consequence, the RASTA filter band passes every function coefficient. The spectral log or the Cepstral domain uses the Rasta filter. In effect, every

coefficient of feature is passed through the RASTA filter band. Using LTI filters, additive components can be easily filtered [4].

Consider the short-term spectrum S (ω, t) and interpreted by the LTI filter with the conversion feature H (ω, t) resulting in the filtered short-term spectrum X (ω, t).

The following equation describes the relationship between S (ω, t), X (ω, t) and H (ω, t).

$$X(\omega, t) = S(\omega, t)\, H(\omega, t) \qquad (4.33)$$

The corresponding log power spectrum of above equation is given as follows,

$$\text{Log}|X(\omega,t)| = \log|S(\omega,t)|^2 \log|H(\omega,t)|^2 \qquad (4.34)$$

The time domain convolution refers to the calculation of the frequency domain and the inclusion of the log power domain.

Thus, the LTI filter can be easily separated if the additive components have different properties in time. .The Rasta PLP method is used to render the PLP study stable against convolution disruptions. Each critical band trajectory log is filtered using a band pass filter (BPF) by compressing the static non-linearity, modifying the BPF characteristics, maximizing recognition accuracy [95].

4.3.1.11 Observations

Two RASTA PLP analytical models are used in this work, namely RASTA PLP spectral analysis and RASTA-PLP cepstral analysis. They are compared on the basis of speech recognition accuracy. The experimental result shows that the RASTA-PLP cepstral analysis provides good recognition accuracy compared to RASTA-PLP spectral.

Observation reveals that between MFCC, LPC, RASTS-PLP spectral and RASTA-PLP cepstral, MFCC feature extraction technique provides better performance compared to others, it has been found that the differences in MFCC-related parameters also affect the results already discussed in section 4.3.1. Therefore, it is

agreed in this study to use MFCC as a technique for the extraction of features with previously defined parameters.

4.4 Implementation of Different classifier

Recognition of speech is an important task in the processing of speech, which is the next stage of extraction of features. This can be achieved using a variety of matching techniques. Classifiers plays important role in determining closet match from pattern matching. It is the problem of defining which of a set of classes a new observation belongs to, based on a training set of data containing observations (or instances) whose category membership is identified. After an optimal subset of features is selected, the classifier can be designed using a variety of approaches.

The algorithm enforcing classification is known as a classifier, especially in the context of a concrete implementation.

Types of Classifiers Used in this Research

1. Minimum Euclidean Distance
2. Support Vector Machine (SVM)
3. K- Nearest Neighbor Classifier (k-NN)
4. Hidden Markov Model (HMM)

4.4.1 Algorithm to find Minimum Euclidean Distance

Function vector series { x1, x2 .. xi) describes an unidentified speech during the speech recognition process, and then being compared to the database codebooks.

This can be done in order to define the unknown speech by measuring the distance of two vector sets based on decreasing the Euclidean distance.

Using Euclidean distance equation, the minimum distance can be determined.

The Euclidean distance between two points P = (p1, p2…pn) and Q = (q1, q2...qn).

$$= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^{\ 2}} \qquad (4.35)$$

The lowest distance speech is chosen as the unknown voice.

4.4.2 Support Vector Machine Algorithm

SVM uses supervised learning techniques to classify data. For both linear and non-linear classification, SVMs can be used. For linear SVMs feature vectors denoted by$S_i \in R^d, i = 1..M$ where M is number of training samples, $d$is the number of speech signal features.

For linear SVMs feature vectors denoted by$S_i \in R^d, i = 1..M$, where $M$is number of training samples, $d$is number of features of speech signal.

To model accurate voice, the speech module classifies the hyperplane in the two categories $O_i = -1$ (Out of training dataset) or $O_i = +1$ (from training dataset).

This hyperplane is defined as follows,

$$w \,.\, S + b = 0 \qquad (4.36)$$

Where, w is normal to the plane, and b is a bias.

SVMs construct a hyperplane using the following formulas to determine the boundary of both groups.

$$w \,.\, S_i + b \geq +1 \; for \; O_i = +1 \qquad (4.37)$$

$$w \,.\, S_i + b \leq -1 \; for \; O_i = -1 \qquad (4.38)$$

The optimal hyperplane can be sought by optimizing the limit margin as per equation (4.37) and (4.38).

The SVMs are applied directly to a higher-dimensional feature space S_i instead of input space R^d in the case of non-linear classification.

$$\Phi : R^d \rightarrow S_i \qquad (4.39)$$

This transformation is done using different types of kernel functions along with equations given in following table 4.4

Table: 4.4 Different kernels in SVM

Type of Kernel	**Equation**
Linear Kernel: $K(S_i, S_j)$	$S_i^T S_j$
Radial Bases Kernel:$K(S_i, S_j)$	$e^{-\left(\frac{\|S_i - S_j\|^2}{2\sigma^2}\right)}$
Polynomial Kernel: $K\ (S_i, S_j)$	$(S_i\ .\ S_j + a)^{\ b}$
Sigmoidal Kernel: $K(S_i, S_j)$	$\tanh(aS_i . S_j - b)$

Where, *a a*nd *b* is Kernel's parameters [4, 5].

Different training measurements are processed and used to assess the distance between specimens that are unknown. In this study, the features are selected in sequential forward sequence (SFS) way.

4.4.3 Algorithm to Find Unknown Sample Using K-Nearest Neighbor Classifier

The KNN algorithm's key concept is that given a test sample, some neighbors are used in KNN to measure the neighboring degrees of testing and training samples on training sets, and then mark them with their nearest neighbor's K tag, if there are a number of labels in their nearest neighbor's K, the samples will be allocated to the nearest neighbor's majority category.

The algorithm uses the improved algorithm one to train samples to cluster training samples, make a relatively uniform distribution of training samples and then based on training samples; a new partition clustering-based KNN classification algorithm is implemented to classify the test samples. Change a dynamic K ' settings in the new algorithm in the iteration.

The purpose of this algorithm is to use a method of initializing the cluster center to decide the initial focal point based on a category pre-set number K. From the original document data pick K objects, each object representing the initial of each cluster core. And then the majority of each entity will be allocated to the cluster of greatest similarity based on the similarity size in all initial cluster centers. K closest neighbor method is used to re-classify the document after focusing on all the test documents. Repeat this process until all forms of records are no longer updated.

There are two parameters to be set, one being cluster number K, the other being K ' used during each of the nearest k-neighbor iterations. If the parameter is too small, you can't find enough documents to identify properly; otherwise there would be more neighbors in the distant clusters, so the report would be incorrectly allocated to more distant clusters. Implement a dynamic K ' change in each iteration, setting the nearest adjacent parameter K ' is the number of documents in the smallest class size of the cluster adding one in the current clustering iteration, thereby preventing the unequal classification phenomenon and assigning the report to the correct clusters.

K-NN is a simple algorithm that stores all existing cases and classifies new cases based on similarity measurements (e.g. calculation of distance). To evaluate instances of ' closeness, ' a distance metric is required. The instance is defined by finding the nearest neighbors and selecting the most common group of neighbors. To evaluate the closest neighbor in k-NN, the distance between the unknown sample and all practice samples is determined.

The problem of classification k-NN is described as:

- Consider qualified data set of vector function data and category labels;

$$Train_{Data} = \{[\underline{a}(1), \underline{b}(1)], [\underline{a}(2), \underline{b}(2)], \dots\dots\dots\dots\dots [\underline{a}(n), \underline{b}(n)]\},$$

- Where $\underline{a}(j)$ denotes the vector feature as j^{th}data
    - $\underline{a}(j)$ Represents j^{th} row of a $n \times c$ matrix, where c corresponds to the MFCC coefficient.
    - $\underline{b}(j)$ indicates class label of the j^{th} feature vector
    - b=1, b=2.., b=n indicates different class values
    - Consider $\underline{z}$ as unknown feature vector
    - Find class label for this unknown feature vector $\underline{z}$
- Search $Train_{Data}$ for the closest feature vector to $\underline{z}$
    - let this “closest feature vector” be $\underline{a}(j)$
- Classify z as $\underline{a}(j)$ mark
    - $\underline{z}$ is assigned a label $\underline{b}(j)$
- To find the nearest neighbor to z from $Train_{Data}$

- Arrange the function vectors in ascending order according to the distance metric. Assign k vectors that give z the nearest distance
- Prediction
    - K vector arrangement provides a set of category labels. Choose from the set the most common class tag('vote')
    - Predict the class of $\underline{z}$ accordingly.

4.4.4 Observations

In this analysis, Euclidean distance measurement is found to provide better average precision for all k values than other distance measurements. However, Hamming and Jaccard's distance measurements do the worst for the database of discourse disorder. The value k also influences the classifier's output and it was found not to be too small or too high.

4.4.5 Algorithm to Predict the Correct Word Using HMM Classifier

In HMM we have two types of states, one state is known as set of hidden state say ω and other state is visible state say v. Let us assume that there are three hidden states, ω is set of $\{\omega_1, \omega_2, \omega_3\}$ and visible state v is set of $\{v_1, v_2, v_3\}$ which is emitted by HMM. Consider θ is hidden Markov model. State representation of basic HMM Model shown in figure 4.14.

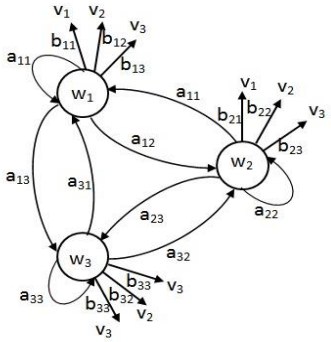


Figure 4.14: State representation of Basic HMM Model

Machine can make the transition from $\omega_1$toω_2 , $\omega_2$to ω_3. Let HMM be in hidden state $\omega_i(t-1) \rightarrow \omega_j(t)$and the probability of this transition is denoted as a_{ij}.

In every state the machine can emit one visible state with the probability of visible state $P(v_k|\,\omega_j) = b_{jk}$ from different hidden states.

$$\sum_{j} a_{ij} = 1 , \quad \forall i \qquad (4.41)$$

Similarly machine always emits visible state

$$\sum_{k} b_{jk} = 1 , \quad \forall j \qquad (4.42)$$

HMM has specific hidden state which is called as receiving state, an accepting state or final state. Once the machine reaches that state it cannot come out from that state. All the transformations are within that accepting state and it will emit only one visible state. Modified HMM state diagram is shown in figure 4.15.

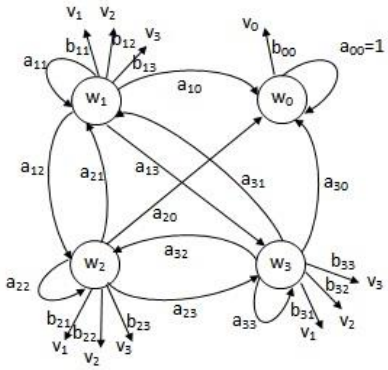


Figure 4.15: Modified HMM state diagram

Incorporating that accept state ω_0 the HMM model is redrawn. From any of the states$\omega_1, \omega_2, \omega_3$ there is transition to $\omega_0$but reverse case is not allowed. It can have the transition within a state say $a_{00} = 1$ and it has only one emission state $v_0$and $P(v_0) = b_{00}$ is also equal to 1. This can be used for temporal analysis [5].

Central issues in HMM

1. Evaluation problem
2. Decoding problem
3. Classifier learning on training problem

1. Evaluation Problem

HMM model is denoted as θ which is characterized by, $\theta \rightarrow \omega, v, a_{ij}, b_{jk}$ and sequence of visible symbols v^T of lengthT. It is important to find $P(v^T|\theta)$. Considered different number of sequences (n) and unknown sequence is to be classified then for every

sequence. Let us build HMM model say $\theta = \{\theta_1, \theta_2, \theta_3, \ldots \ldots \theta_n\}$ that means find out visible sequence v^Tprovided to know the model of θ.

2. Decoding

The most probable sequence ω^T of hidden states that led to the creation of v_0 to find out the sequence ω^T has created the sequence v^T called the decoding problem.

3. Classifier learning on training problem

Learning problem is given rough structure of the HMM that is how many hidden state and visible structure the hidden Markov model has?

$$\omega, v \rightarrow \text{known} \qquad (4.43)$$

From given training set estimate transition probabilitya_{ij}, emission probability b_{jk} and How to evaluate $P(v^T|\theta)$.

$$P(v^T|\theta) = \sum_{r=1}^{r_{max}} P(v^T|\omega_r^T)P(\omega_r^T) \qquad (4.44)$$

Where,T is length of visible sequence and index r indicates one of the possible sequences and, $\omega_r^T = \{\omega(1),\ \omega(2)\omega(T)\}$

Let N be the number of hidden states and define $r_{max} = N^T$

$$P(\omega_r^T) = \prod_{t=1}^{T} P\big(\omega(t)\big|\omega(t-1)\big) \qquad (4.45)$$

These are transition probability and$P(v^T|\omega_r^T)$ probability of emission of visible symbol.

$$P\Big((v^T|\omega_r^T)\Big) = \prod_{t=1}^{T} P\big(v(t)\big|\omega(t)\big) \qquad (4.46)$$

$$(v^T|\theta) = \sum_{r=1}^{r_{max}} \prod_{t=1}^{T} P(v(t)|\omega(t))P(\omega(t)|\omega(t-1)) \qquad (4.47)$$

To solve this equation (4.46) very large computation is needed. The complexity of the equation will be of the order $O(N^T.T)$. Instead of that we can use recursive algorithm. In the recursive algorithm $\propto_j (t)$is defined as the probability that the machine will be

in the state ω_j in time step t after emitting first t number of visible symbols in the sequence of v^T.

$$\propto_j (t) = \begin{cases} 0 & t = 0 \text{ and } j \neq \text{initial state} \\ 1 & t = 0 \text{ and } j = \text{initial state} \\ \left[\sum \alpha_i(t-1).a_{ij}\right].b_{jkv(t)} & \text{otherwise,} \end{cases} \quad (4.48)$$

Where $b_{jkv(t)}$ is emission probability in state ω_j, and a_{ij} is transition probability.

4.4.6 Observations

Based on phonemes, the HMM model is designed. In the case of normal speech, all-digit phonemes are present and follow a sequence for a word from left to right. But the same phonemes are missing in the case of disordered speech, i.e. few temporal characteristics are not present, and as a result the accuracy of recognition for disordered speech is less than normal speech.

4.4.7 Algorithms used in Different Stages of HMM

A. Evaluation

The following algorithm is used in evaluation stage.

- Forward recursive algorithm (Dynamic Programming)
  1. Initialize :$t \leftarrow 0, a_{ij}, b_{jk}, v^T$ and $\propto_j (0)$
  2. For $t \leftarrow t + 1$

$$\propto_j (t) = b_{jkv(t)}.\sum_{i=1}^{N} \alpha_i(t-1).a_{ij} \quad (4.49)$$

  where, N is number of hidden states

  3. Until t=T
  4. Return $P(v^T|\theta) \leftarrow \propto_0 (T)$ for final state
  5. Stop

  By using this algorithm find out what is the probability of state in every hidden state. How the forward algorithm works in HMM is shown in figure 4.16.

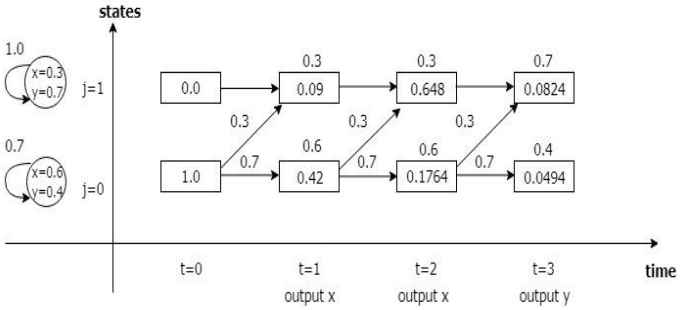


Figure 4.16: Working of forward algorithm

This computes the probability of output sequence $v_1^3 = (x, x, y)$could have been generated by HMM.

The calculation of process starts from the first state to the last state within a time frame before moving to the next time frame. The last cell produces the probability of sequence (x, x, y) is 0.082.

B. Decoding

In this stage, find the most likely sequence of hidden states or paths using the given length of the visible state v^T.

- **Algorithm**

1. Initialize: Path $\leftarrow \{\ \}$; $t \leftarrow 0, j \leftarrow 0$.
2. For $t \leftarrow t + 1$

$$j \leftarrow j + 1$$

3. For $j \leftarrow j + 1$

$$\propto_j (t) = b_{jkv(t)} . \sum_{i=1}^{N} \alpha_i (t-1) . a_{ij} \qquad (4.50)$$

4. Until $j = N$

$$j' \leftarrow \text{argmax}\ \alpha_j(t)$$

5. Append ω_j, to path
6. Until $t = T$
7. Return Path

C. Learning

$\beta_i(t)$ is the probability that the model will be in $\omega_i(t)$ and will generate remainder of the given target sequence v^T.

$$\beta_i(t) = \begin{cases} 0 & \omega_i(t) \neq \omega_0 \text{ and } t = T \\ 1 & \omega_i(t) = \omega_0 \text{ and } t = T \\ \left[\sum \beta_j(t+1).a_{ij}\right].b_{jkv(t+1)} & \text{otherwise} \end{cases} \quad (4.51)$$

Where,ω_0 is the last hidden state, and T is the last symbol in sequence of state.

The backward algorithm says about the probability that the machine or the model will be in state ω_j at time step t will generate the remaining part of state of visible symbol.

- **The backward algorithm**

1. Initialize :$t \leftarrow T, a_{ij}, b_{jk}, v^T$ and $\beta(T)$
2. For $t \leftarrow t-1$

$$\beta_i(t) = \sum_j \beta_j(t+1).a_{ij}.b_{jkv(t)} \quad (4.52)$$

3. Until t=1
4. Return $P(v^T) \leftarrow \beta_i(0)$ for the known initial state
5. End

So estimating the machine's likelihood in the state ω_i or ω_j in the time step t produces the remaining symbols from the given state string.

Forward-backward algorithm (Baum welch) will be used for correct estimation for the values of transition probability a_{ij} and emission probability b_{jk}.

As the number of hidden states and visible states is known, consider a_{ij} and b_{jk} correct value.

Correct estimation of transition probability a_{ij} and emission probability b_{jk} is needed.

- **The forward backward algorithm (Baum welch)**

1. Initially assume random values of $a_{ij}$and b_{jk}.
2. Using forward algorithm estimate $\propto_j (t)$ as,

$$\propto_j (t) = b_{jkv(t)}. \sum_{i=1}^{N} \alpha_i(t-1). a_{ij} \qquad (4.53)$$

3. Using backward algorithm estimate $\beta_j(t)$ as,

$$\beta_i(t) = \sum_j \beta_j(t+1). a_{ij}. b_{jkv(t)} \qquad (4.54)$$

4. The equations from step 2 and 3 uses a_{ij} and b_{jk} but the challenge is to estimating a_{ij} and b_{jk} is nothing but the learning algorithm. These estimates are not exact values of a_{ij} and b_{jk}, but they are only approximations. So to find out refined values of a_{ij} and b_{jk} , define a probabilities of the transition state $\omega_i(t-1)$ to $\omega_j(t)$ in the time step t to $t-1$ for specific tuning sequence v^T. Let us say this probability is denoted as $\gamma_{ij}(t)$ and defined as,

$$\gamma_{ij}(t) = \frac{\alpha_i(t-1)a_{ij}b_{jk}\,\beta_i(t)}{P(v^T|\theta)} \qquad (4.55)$$

5. Find refined values of a_{ij} and b_{jk} by using following equation.

   Expected number of transition $\omega_i(t-1) \rightarrow \omega_j(t)$ at any time of sequence v^T.

$$\sum_{t=1}^{T} \gamma_{ij}(t) \qquad (4.56)$$

Total expected number of transition from ω_i to any state is equal to

$$\sum_{t=1}^{T} \sum_{k} \gamma_{ik}(t) \qquad (4.57)$$

After knowing these two values then calculate refined value of $\hat{a}_{ij}$ and$\hat{b}_{jk}$ by using,

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T} \gamma_{ij}(t)}{\sum_{t=1}^{T} \sum_{k} \gamma_{ik}(t)} \qquad (4.58)$$

$$\hat{b}_{jk} = \frac{\sum_{t=1}^{T} \sum_{l} \gamma_{jl}(t)}{\sum_{t=1}^{T} \sum_{l} \gamma_{jl}(t)} \qquad (4.59)$$

The numerator satisfies the condition $v(t) = v_k$ and denominator is irrespective of v_k

6. Repeat the above steps number of times unless the change in a_{ij} and b_{jk} is within tolerable state. Then these values can be used to find evaluation process. In this way the HMM is trained.

4.4.8 Observations

Based on phonemes, the HMM model is designed. In the case of normal speech, all-digit phonemes are present and follow a sequence for a word from left to right. But the same phonemes are missing in the case of disordered speech, i.e. few temporal characteristics are not present, and as a result the accuracy of recognition for disordered speech is less than normal speech.

4.4.9 Algorithm to recognize correct word Using ANN

An ANN can works the two facts such as,

- By choosing the correct network topology and weight values, measure any computable function.
- It is learning from observation, through trial-and-error in particular.

ANN is made up of various types of neural networks feed-forward networks and feed-back networks. Feed-forward NNs allow only one way for signals to travel; from input to output. There is no input (loops), i.e. no layer output affects the same layer. Feed-forward NNs tend to be direct networks associating inputs with outputs. They are commonly used to recognize patterns. This kind of management is also appealed to as bottom-up or top-down.

Feedback networks may have signals that propagate in both directions through the addition of loops on the network. Feedback networks are dynamic; their ' state ' is constantly changing until they reach a point of balance. They remain at the point of equilibrium until the input changes and finding a new equilibrium. Feedback architectures are also referred to as dynamic or recurrent, although in single-layer organizations the latter term is often used to describe feedback connections.

- **Levenberg-Marquardt algorithm**

It is also known as the least square damped approach designed specifically to deal with loss functions that take the form of a number of square errors**.** This functions without the exact Hessian matrix being determined. Instead, this works with the Jacobian matrix and the gradient vector.

Suppose f: $\mathbb{R}^n \to \mathbb{R}$ is a function which inputs a vector $\mathbf{x} \in \mathbb{R}^n$ and outputs a scalar $f(\mathbf{x}) \in \mathbb{R}$; if all second partial derivatives of f exist and are continuous over the function domain, then the Hessian matrix H of f is a square $n \times n$ matrix, usually defined and arranged as follows:

$$\mathrm{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

$\mathrm{H}(\mathrm{f}(\mathrm{x})) = \mathrm{J}(\mathrm{f}(\mathrm{x}))\mathrm{T}$, can be considered to be related to the Jacobian matrix. In vector calculus, the Jacobian matrix is the matrix of all of a vector-valued function's first-order partial derivatives.

- Consider the loss function that can be expressed as the sum of squared errors of the form.

$$f = \sum e_i^2 , \quad i = 2, \ldots\ldots, m \tag{4.43}$$

Here *m* is the number of instances in the data set.

- Define the Jacobian loss matrix as containing the error derivatives with respect to the parameters;

$$J_{i,j} f(w) = de_i / dw_j (i = 1 \ldots, m \,\&\, j = 1, \ldots, n) \tag{4.44}$$

Where, m is the number of instances in the data set and n is the number of neural network parameters. Remember that the Jacobian matrix size is m×n approximately.

- The loss function's gradient vector can be calculated as:

$$\nabla f = 2 J^T . e \tag{4.45}$$

Here *e* is the vector for all terms of error.

- Approximate the Hessian matrix with the following expression

$$Hf \approx 2 J^T . J + \lambda I \tag{4.46}$$

Where λ is a damping factor that ensures the positiveness of the Hessian and I is the identity matrix.

- Define the parameters improvement process with the Levenberg-Marquardt algorithm

$$W_{i+1} = W_j - \left(j_j^T \,.\, J_i + \lambda_i I\right)^{-1} (2J_i^T . e_i) \qquad i = 0,\ 1,\ \ldots \quad (4.47)$$

- Adjust the value λ
- After calculating the loss, the gradient and the Hessian approximation. Then the damping parameter is adjusted so as to reduce the loss at each iteration.

4.4.10 Observations

It was found that the accuracy is low compared to other classifiers and depends on neural network transmission. For very large data sets and neural networks, the Jacobian matrix is becoming vast and therefore requires a lot of memory.

In section 4.2, different algorithms have been developed and implemented for the extraction of attributes and speech recognition techniques. After applying different algorithms to the dataset, the experimental result shows that the MFCC as a feature extraction technique and the k-NN classifier provides better accuracy among all other techniques. Therefore, holding these approaches same along with the same parameters, the MFCC algorithm for feature extraction and k-NN classifier was used for speech rectification with phoneme separation.

4.5 Proposed system

The proposed system is designed to recognize the correct word for articulatory handicapped people and also to rectify the corrected word. In this system, the speech signal (digits zero to ten) was segmented into phonemes after speech acquisition. Speech segmentation is important step of this algorithm.

4.5.1 Speech segmentation

Speech segmentation splits continuous waves of sound into some basic units, such as words, phonemes or syllables, which can be understood. Segmentation may also be used to distinguish between different forms of audio signals and large amounts of audio content, also referred to as audio sorting. Automatic voice segmentation methods can be categorized in many ways, but the division of blind and assisted segmentation algorithms is a very specific category.

- **Blind segmentation**

The word blind segmentation applies to approaches where there is no pre-existing or current understanding of the linguistic properties of the signal to be segmented, such as orthography or complete phonetic annotation.

- **Assisted segmentation**

Assisted segmentation algorithms use some sort of external speech stream linguistic information to segment it into the correct segments of the desired kind. Orthographic or phonetic transcription is used as a parallel reference to expression, or as a learning method for such data [15].

In the fig.4.12 describes the steps related proposed system of speech recognition for articulatory handicapped people with phoneme separation. It is a segmentation of the speech process. This study uses an assisted segmentation algorithm. The features are extracted by using the MFCC technique and maintaining the same parameters of the MFCC as decided in section 4.3.1. Twenty coefficients of each segment are considered. The proposed system is developed for 0 to 10 digits, which contains eighteen phonemes.

Depending on the standard lexicon table shown in Table 4.1[96], the digit is divided into 2, 3, 4 and 5 segments respectively. Out of 1100 samples, 990 samples are used for system training. The proposed system is developed for 0 to 10 digits, which contains eighteen phonemes. Therefore feature extraction matrix contains 3150 rows and 20 columns. Labels are assigned for each class of phoneme. K-NN classifier is used to classify the phoneme.

This will be used as reference data for comparing segments of tested data and finding the correct phoneme. Not only to identify the matching phoneme, but also to find the exact location and frequency of the phoneme in order to recognize the correct word, this can be done by using the current algorithm known as the positive position searching algorithm. The word is predicted after finding the position and occurrence of the phoneme, using the technique of string concatenation

4.6 Structure of proposed system

In figure 4.17 proposed system of speech recognition for articulatory handicapped people with phoneme separation is shown.

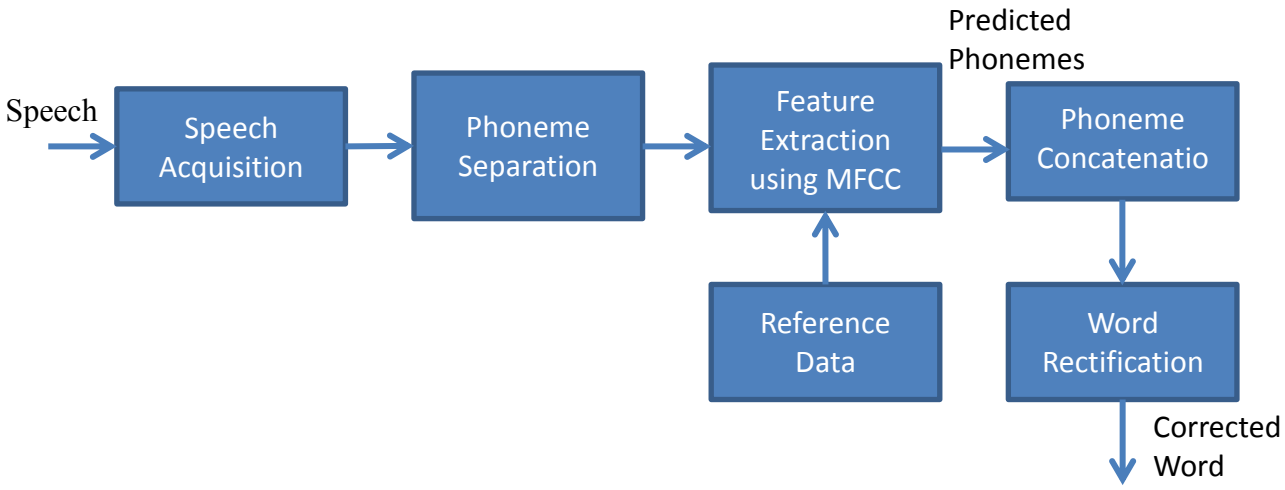
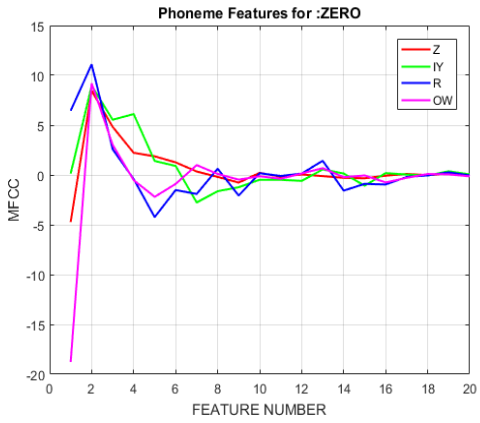


Figure 4.17: Proposed system of speech recognition for articulatory handicapped people with phoneme separation

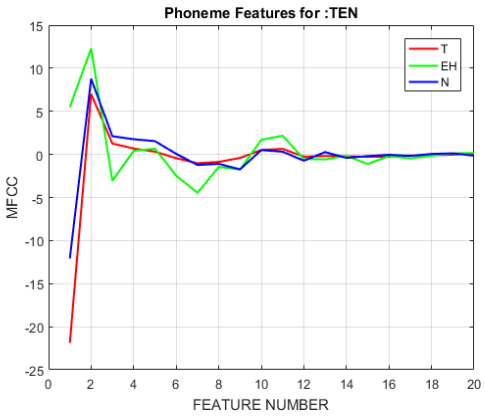
Algorithm for speech recognition for articulatory handicapped people with phoneme separation

- Sample fs=44100Hz speech signal
- Count the total number of samples for each word
- Split each word into 2, 3, 4 and 5 segments using assisted segmentation
- Use MFCC to extract the features for each segment
- Use k-NN classifier to classify the segment with pretrained phoneme dictionary
- Concatenate predicted phonemes for all segments
- Find correlation of predicted phonemes with lexicon table by using position and occurrence base searching (positive position searching)
- Predict the word which has maximum score depending upon correlation & Occurrence

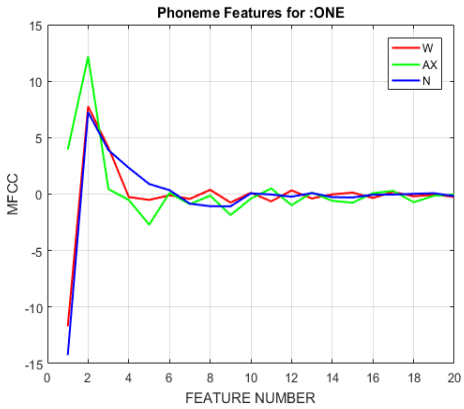
The characteristics of each phoneme are extracted using the MFCC feature extraction technique. The following figure 4.18 (a)-(k) shows distribution of features of each phoneme for each digit.


Phoneme Features for :ZERO
Z
IY
R
OW
MFCC
FEATURE NUMBER

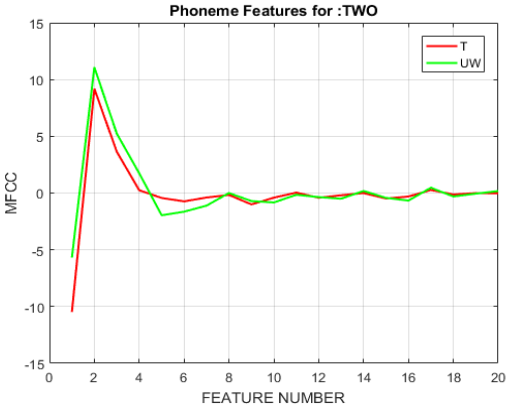
(a)


Phoneme Features for :TEN
T
EH
N
MFCC
FEATURE NUMBER

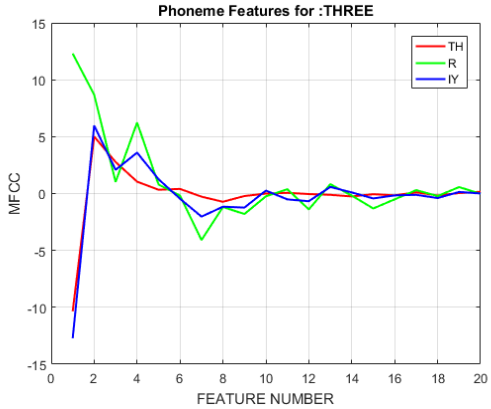
(b)


Phoneme Features for :ONE
W
AX
N
MFCC
FEATURE NUMBER

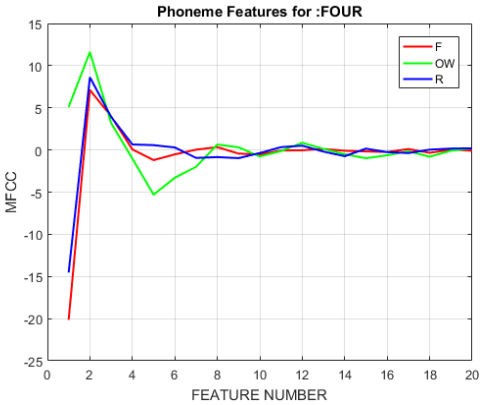
(c)


Phoneme Features for :TWO
T
UW
MFCC
FEATURE NUMBER

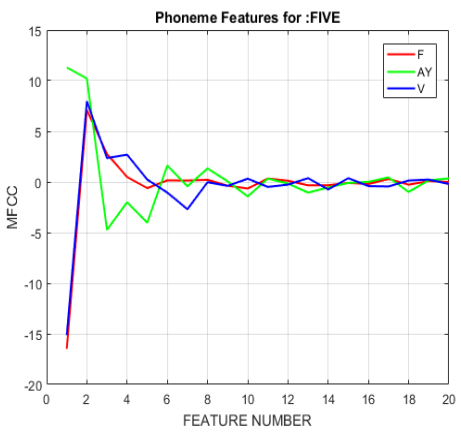
(d)


Phoneme Features for :THREE
TH
R
IY
MFCC
FEATURE NUMBER

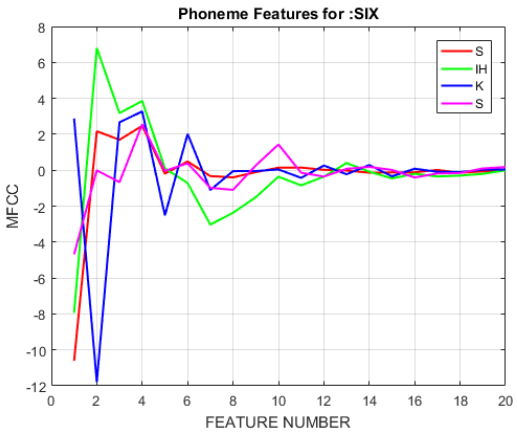
(e)


Phoneme Features for :FOUR
F
OW
R
MFCC
FEATURE NUMBER

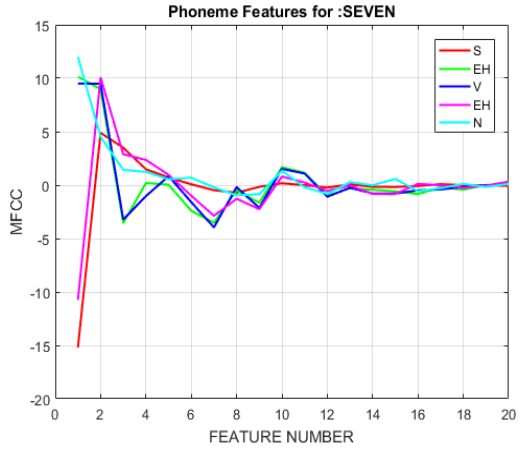
(f)



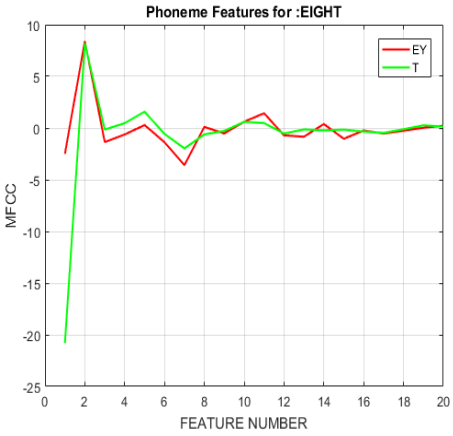
(g)



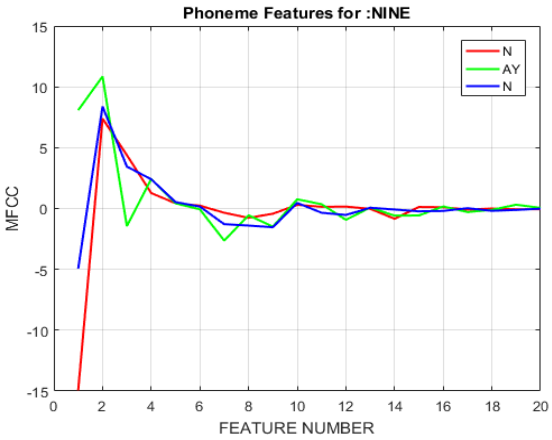
(h)



(i)



(j)



(k)

Figure 4.18: (a)-(k): Value of each coefficient for each phoneme

The following figure 4.19 shows distribution of features of each phoneme for all digits.

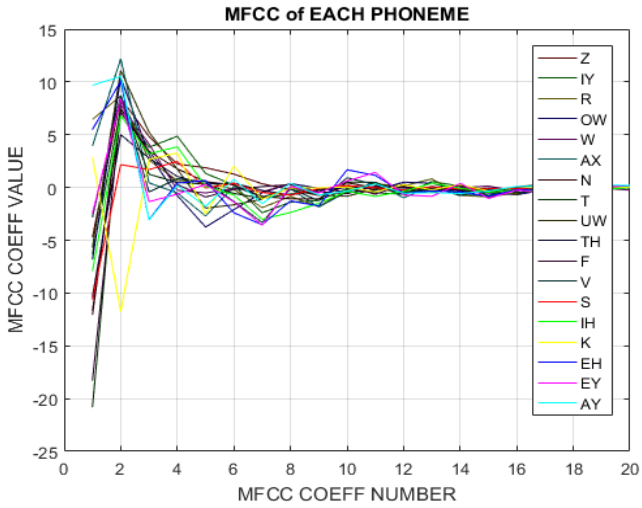


Figure 4.19: Coefficient for each phoneme

4.6.1 Proposed Positive Position Searching Algorithm

Consider a string in the proposed algorithm as a 0-indexed sequence. A string is a character set. In fact, a string S = "ten" is an array [' t ', ' eh ', ' n ']. A string's number of characters is called its length and denotes it by |S|. Consider S[i] as the string character at position i.

A substring is a sequence of a string's consecutive contiguous elements, denote the substring beginning with I and ending with S [i ... j] at j of string S.

The prefix for string S is a substring which begins at position 0, and a suffix is a substring which finishes at |S|-1. The correct S prefix is a prefix that is different from the S prefix. Similarly, the S suffix is a suffix that is different from the S suffix. The + operator must indicate the concatenation of the sequence.

All occurrences of a pattern M are calculated from the given text N. In the following example, all occurrences of M in N are highlighted from the string N and pattern M.

M= six

N=slixx**six**sxxxxhixsslixh**six**sxxiihlshsh**six**xsssshxhs**six**hsxsxhh

One simple way to solve the problem is by iterating all i from 0 to |N| - |M|, and checking if there is an N substrings that begins at i and matches with M:

There are two possibilities to find a match starting at position i on the text.

Case No1: Match is not found

Then there is at least one index that doesn't match the pattern in the text. Let i+j be the smallest of such indices: N [i...i+j-1] = M [0...j-1] and N[i+j] ≠ M[j]

The following figure 4.20 describe the process of string matching function.

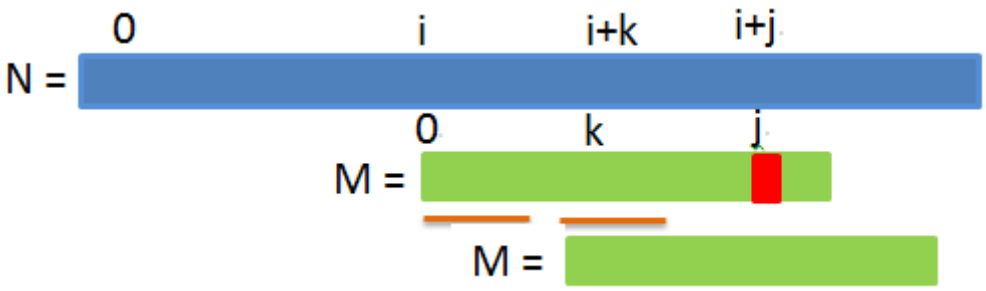


Figure 4.20: String matching function

Above example demonstrates string matching function. Since there is no match at position i it should start from another place to find a match, but the question is where to start from. If it begins from the next i+1 position then the probability is that it might end up finding a mismatch even before i+j-1 in a position. It could at least start from a position that ensures the string matches up to i+j-1. Therefore, seeking a match starting from the smallest i+k will continue to match N [i+k ... i+j-1] with some N.

If N matches M from i to i+j-1, then N [i+k ... i+j-1] is a suffix of M [0 ... j-1]. This means that if the mismatch is found at position j, it should start at the smallest k, so that M [k ...j-1] is a prefix of M. k is the smallest, so M[k ... j-1] is the biggest appropriate suffix, which is also a prefix. From now on, say it is "border" to the proper prefixes that are also proper suffixes (for example, the PQRSPQRSPQ string has two PQRSPQ and PQ boundaries).

Case No2: Match is found

Using the same logic, it should start finding a fit from the smallest k so that M [k ... j-1] is the right prefix for M.

An example of the above two cases described in Table 4.5

Table 4.5: Example related to match found.

	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8
U	F	O	U	R	**F**	**O**	**U**	**R**	**F**	**O**	**U**	**R**	**F**	O	U	R	P	E	N	F	O	U	R	L	K	M	T	N	F
V					F	O	U	R	F	O	U	R	P	E	N															
V									F	O	U	R	F	O	U	R	P	E	N	---------- [MATCH!]									
V																			F	O	U	R	F	O	U	R	P	E	N

The method for identifying occurrences is started from position 4 in table 4.5, but at position 12 there is a mismatch. Before getting a mismatch, the string FOURFOUR has already been matched, and the largest proper prefix that is also a correct suffix (border) of FOURFOUR is FOUR, so it should start finding occurrences from position 8 again and considering that all characters from 8 to 11 are already matched. It turns out that there is an occurrence starting at position 8, so all pattern characters matched, since FOURFOURPEN has no border, then again to find occurrences starting at position 18.

4.6.2 Character concatenation

After finding correct character, concatenation of characters must be done so as to form digit. The procedure is explain as below, which is shown in table 4.6

For 2 segment,

S1 = “T”

S2 = “EH”

S = S1 + S2 = “TEH”

Table 4.6: Phoneme segmentation for testing sample TEN

2 segment	3 segment	4 segment	5 segment	Standard Phoneme
THE	EYNN	TEHVIY	TEHEHNF	TEHN
SHE	TTT	WWEHN	WTEHEHS	TEHN
ZIY	ZUWR	ZZIYR	ZWIYRUW	Z IY R OW
ZR	IYWV	ZIYRN	THIYROWN	Z IY R OW

Similarly for 3, 4 and 5 segment concatenation of string is done. After the concatenation, the phonemes string S is compared with each string of phonemes from training set (2 segment phoneme strings from testing set is compared with 2 segment phoneme string of training set, 3 segment phoneme strings from testing set is compared with 3 segment phoneme string of training set, 4 segment phoneme strings from testing set is compared with 4 segment phoneme string of training set and 5 segment phoneme strings from testing set is compared with 5 segment phoneme string of training set).

The match score of the phoneme string is calculated by finding correlation and occurrences between the phoneme from training and testing of each segment. Afterward, word prediction is done by selecting maximum score of correlation and occurrence of phoneme string. The phoneme string which gives maximum score is correctly predicted.

4.6.3 Observations

Performance of speech recognition and rectification for articulatory handicapped people using phoneme separation shall be observed. It is found that the device output is enhanced with this proposed algorithm compared to speech recognition without separation of phonemes. It was observed that MFCC provided better accuracy of recognition than others compared to other feature extraction techniques such as LPC, MFCC and RASTA-PLP. MFCC output is observed by varying different parameters such as pre-emphasis filter order, frame width, frame overlap percentage, type of window used, and number of coefficients used. These parameters are selected on the basis of the accuracy of the recognition.

After selecting the appropriate extraction technique, the system performance was observed using different classifiers, such as the minimum distance classifier, HMM, SVM, ANN and k-NN classifiers. It has been observed that, among all these classifiers, the k-NN classifier provided better performance of the system without the separation of the phoneme. Thus, the performance of the proposed system has been observed, maintaining these two techniques in the same way. It has been observed that the performance of the system is improved by the use of these two techniques.

It has also been observed how the accuracy of the system is affected by the number of speech segments. The performance of the proposed system is good by using assisted segmentation. The performance of the system was checked for different segments, such as 2 segments, 3 segments, 4 segments and 5 segments. It has been observed that the results for five segments are the most promising. The correction of speech is made on the basis of a positive positioning search algorithm. In addition, the performance of the classifier was also observed with the help of a confusion matrix using parameters such as error rate, accuracy, sensitivity, specificity and F-score. Reports were explored in more depth in Chapter 5.

Chapter 5

RESULTS AND DISCUSSIONS

In the following points, the major contribution of the research work was summarized. The research contribution was divided into two parts, with and without separation techniques for phonemes. Using MATLAB 2017b software, all algorithms are implemented. The research analysis was carried out with the aid of a database composed of 1100 specimens. Out of 1100 samples, 990 samples are used for system training and 220 samples are used for system testing. Dataset consists of utterances spoken by different speakers suffering from an issue of articulatory dysfunction.

5.1 Without phoneme separation

This study uses different methods of extraction of features, such as Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Coding (LPC), and Relative spectral perceptual linear prediction (RASTA-PLP). The results of feature extraction methods will be discussed in the following section.

5.1.1 Mel Frequency Cepstral Coefficient (MFCC) as a feature extraction method

MFCC's findings were analyzed for the following criteria;

- Order of pre-emphasis filter
- Frame size
- Overlapping of window
- Type of window
- No. of filters used

In order to finalize the parameters of the MFCC feature extraction technique, performance of the system was observed by varying parameter values.

I. Pre-emphasis filter:-

- The pre-emphasis 1st order FIR high pass filter was designed with coefficient as a=0.95 and implemented.
- The purpose of pre-emphasis filter was to compensate the high-frequency part.

- Table 5.1 together with Figure 5.1 shows the pre-emphasis high-pass FIR filter effect.

Table 5.1: Effect of order of Pre-emphasis filter

Window Type	1st order, Fs	2nd order, Fs	1st order, Fs/2	2nd order, Fs/2
Ham	67	67	60	58
Han	65	65	62	62
Rect	63	50	61	49
Tri	64	64	58	58

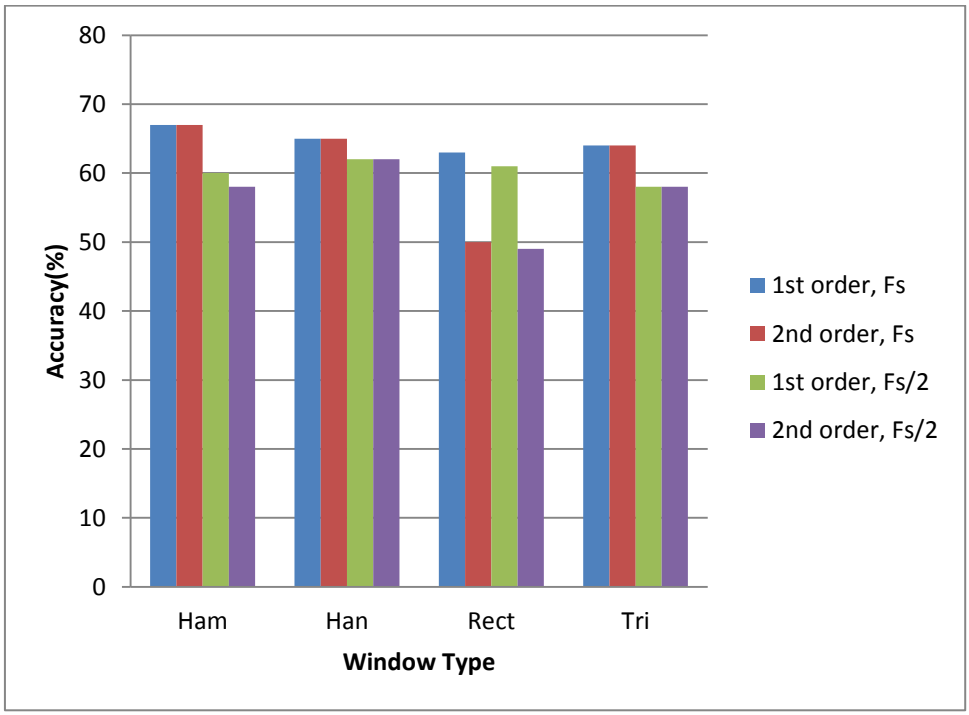


Figure 5.1: Selection of order of pre-emphasis filter

The result shows that the change in order of the FIR high-pass filter does not affect the accuracy of the recognition. Therefore, a FIR high pass filter is used in the proposed 1st order system.

- The graph below, shown in Figure 5.2, describes the determination of the coefficient of ' a ' of the pre-emphasis filter. The result shows that the accuracy of speech recognition is greater than ' a ' =0.95.

- 'a ' = 0.97 attenuate low frequencies more than ' a ' =0.9.
- The choice for value ’a’ depends on the nature of the medium or channel that will be used to store or convey the message signal. The effect of filter coefficient 'a ' describe in figure 5.2

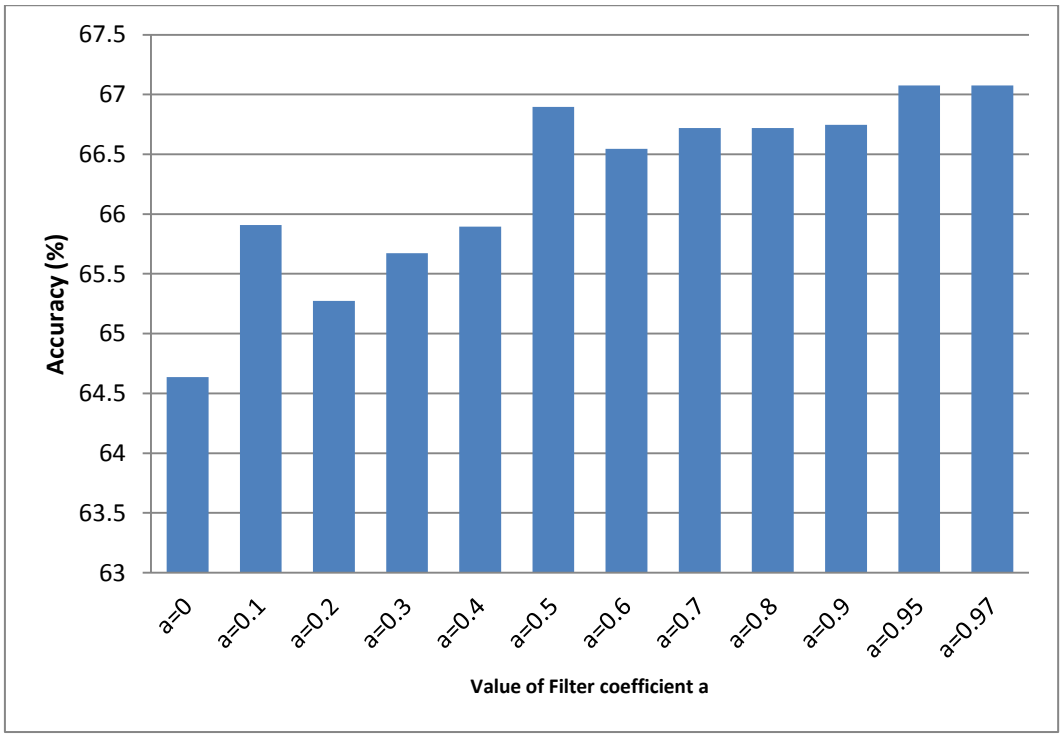


Figure 5.2: Selection of Pre-emphasis coefficient factor ‘a’

I. Frame size, overlapping of frames and windowing:

Speech signal is not stationary in nature, so it is not possible to perform a Fourier analysis, but during a short period of time the signal behaves as stationary and therefore traditional computational methods can be used for analysis purposes. So the framing of the signal is needed. In order to ensure continuity between the two frames, it is necessary to overlap the frames. Frame size and frame overlap are key factors that play a key role in the performance of the feature extraction method. The size of the frame should not be too large or too small. If the size of the window increases, the speech segment analyzed may become non-stationary. If the length of the analysis frame is too short, some signal characteristics will be lost. In

addition, the shortening of the window size shorter than the pitch period (2–3 ms) may fail to register the pitch peaks [61].

Another important aspect of the study is the degree of overlapping or shifting of the research window (frame). The size of the window shift determines the specificity of the speech dynamics data. The lower frame shift price we use, the more information we can get about speech dynamics. Such an assessment can take longer, however, and does not necessarily mean a higher rate of speech recognition. The overlap is usually chosen equal to half or one third of the size of the study window [62], [64]. Nevertheless, some original ideas for setting the frame shift (or frame rate) were suggested. The approach for the selection of frame rates based on a posteriori signal-to-noise ratio is proposed in [65]. The approach is capable of assigning a higher frame rate to a rapidly changing speech and a lower frame rate to a more steady-state speech. Another idea is to adjust the frame shift according to phonetic data [63]. The graph below shows that the variation in window size versus accuracy of recognition by changing window overlaps. In this approach, frame size of 25ms with 60 percent overlapping is used to implement the MFCC algorithm because it gives more accuracy than the different frame sizes shown in Figure 5.3.

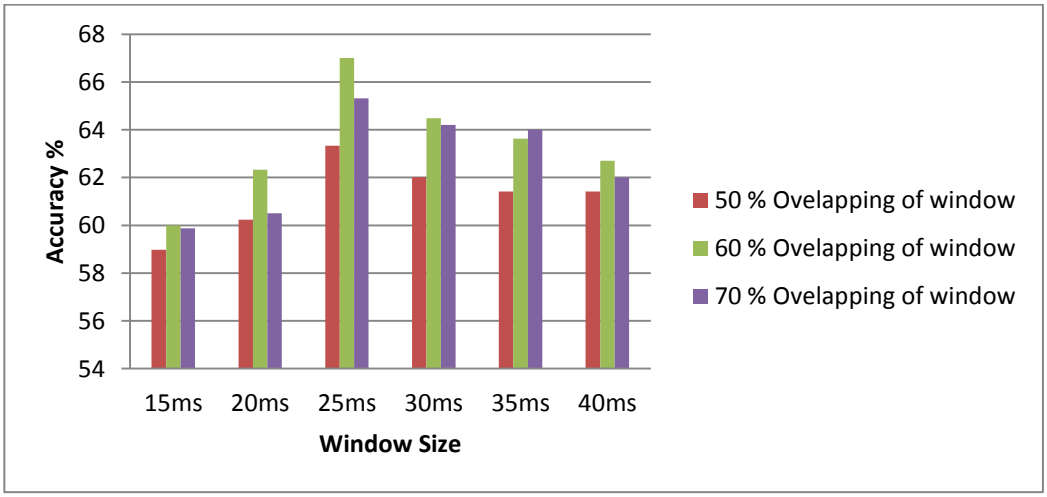


Figure 5.3: Selection of frame size and overlapping of frames

II. Window type

- The simulated output was analyzed for different types of windows such as rectangular, triangular, hanning and hamming. Following figure 5.4 shows the relation between type of window and accuracy. In this research, that hamming window provides better result than other type of window.

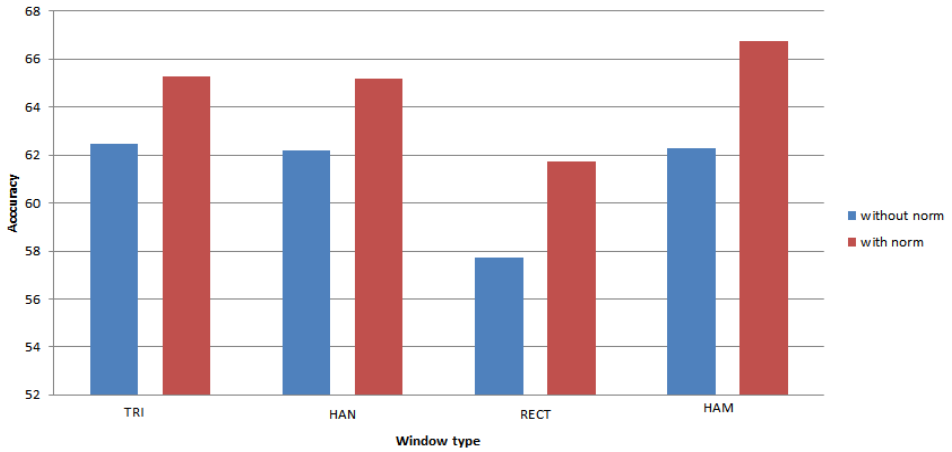


Figure 5.4: Selection of Window Type

Frame windowing is used to reduce the spectral effect and smooth the FFT computing signals. Hamming window main lobe size greater than rectangular window but the same as hamming window, and hamming window's main side lobe attenuation is higher. Blackman window has more side lobe attenuation than any other type of window, but it has larger main lobe width so the transition band width is more affecting system accuracy. The window type is also selected on the basis of factors such as Sidelobe cancellation, Worst Case Processing Loss (WCPL) and Equivalent Noise Bandwidth (ENB).The following figures 5.5, 5.6 and 5.7 display the sidelobe, WCPL and ENB output comparison of window function.

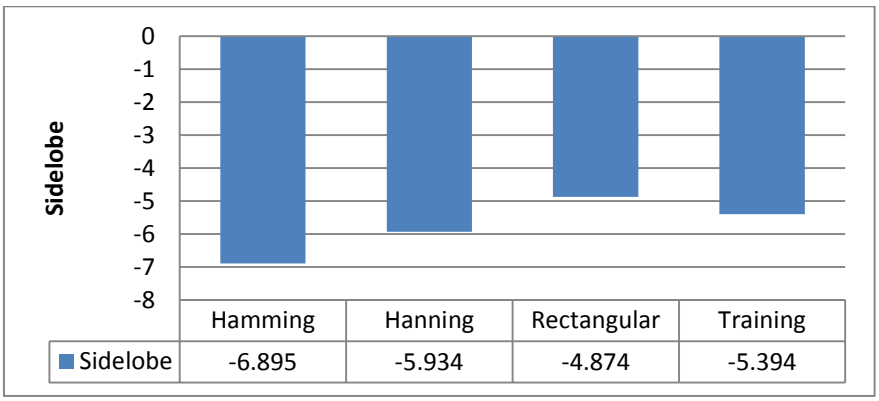


Figure 5.5: Selection of Windowing Function based on Sidelobe Cancellation

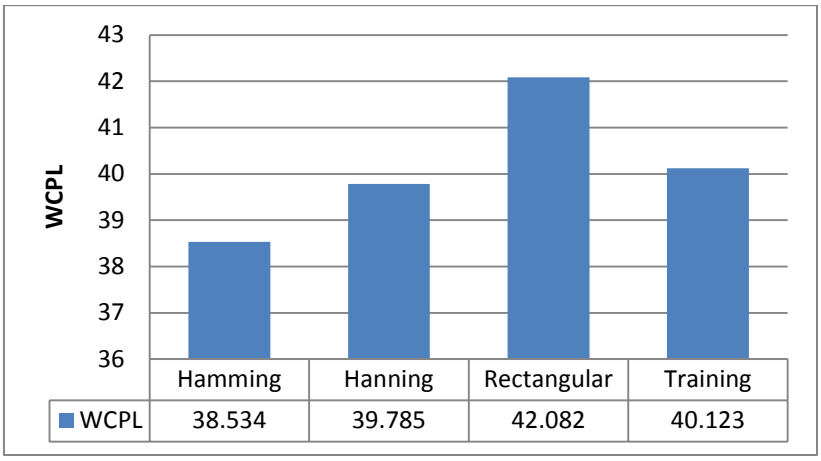


Figure 5.6: Selection of Windowing Function based on WCL Cancellation

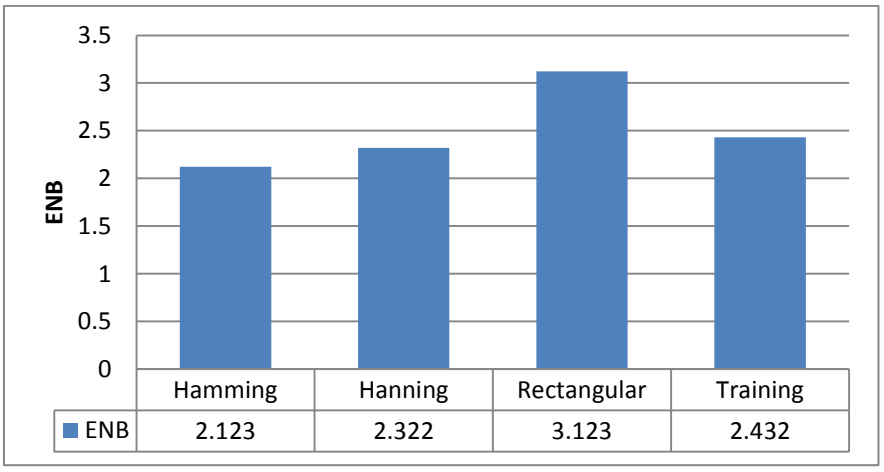


Figure 5.7: Selection of Windowing Function based on ENB Cancellation

The experimental results show that limited sidelobe, fewer WCPL and better ENB values are given by the hamming window.

The hamming window is therefore used in the proposed system to window a segmented speech signal. Then the resulting windowed signal is given for the feature extraction techniques.

III. Number of cepstral coefficients, FFT size

A Cepstral coefficient of MFCC, and FFT size also plays an important role in determining accuracy of system. It is shown in Table 5.2, Figure 5.8 and Figure 5.9.

Table 5.2: Effect of number of cepstral coefficients

Window type	c8	c12	c16	c20	c24	c28	c32	c36	c40
Hamm	58.45	60.63	63.27	67.18	67.36	67.08	67.18	66.95	66.45
Trian	58.27	60.09	62.45	66.18	66.18	66.18	66.18	66.72	65.18
Hann	58.36	60.81	62.18	66.54	66.54	67.09	65.18	66.81	65.81
Recta	56.90	57.36	61.63	65.45	66.72	66.81	65.91	66.09	65.81

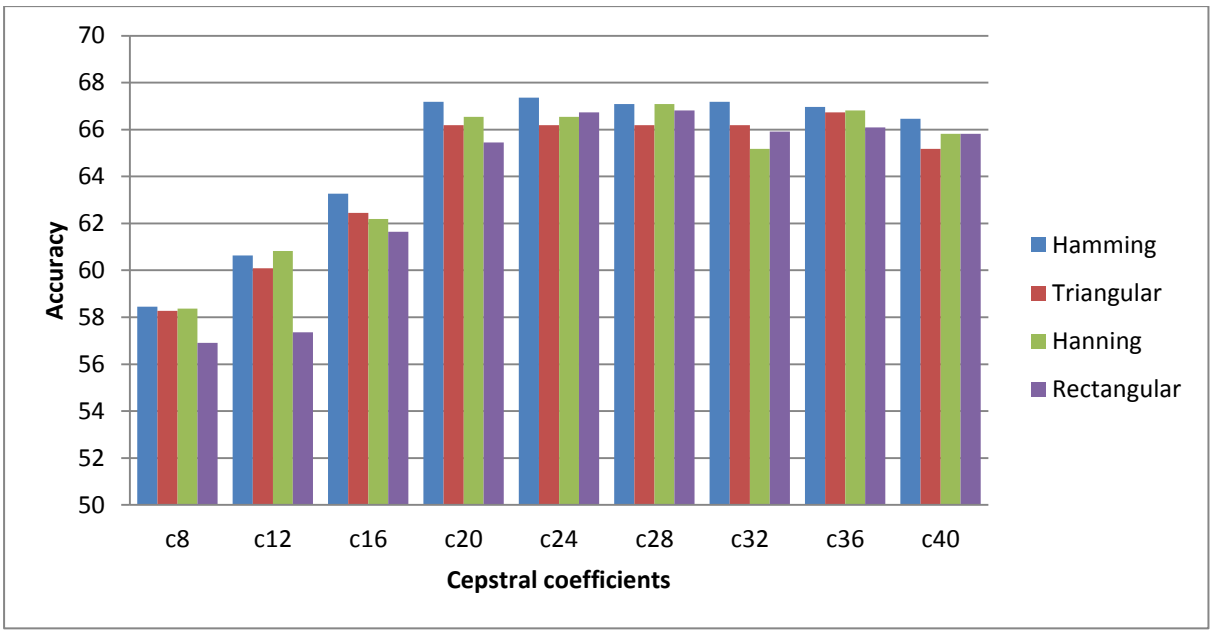


Figure 5.8: Effect of number of Cepstral coefficients

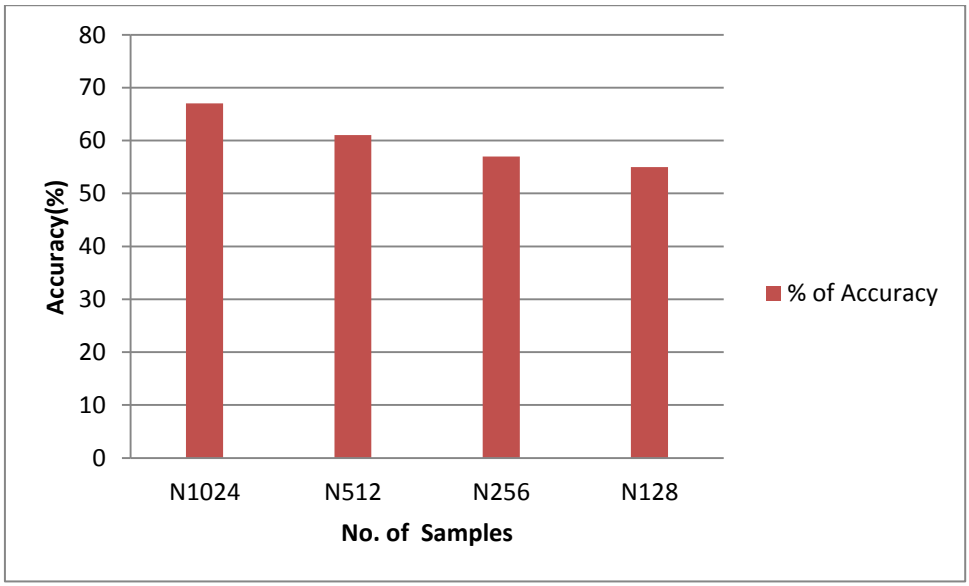


Figure 5.9: Effect of Number of samples per FFT frame

Experimental result shows that twenty cepstral coefficients for the FFT size of 1024 samples with 60 percent overlapping of frames provide better accuracy in this analysis.

Since speech accuracy is maximum for a frame duration of 25 ms, the frame duration in number of samples for a sampling frequency of 44.1 KHz signal is 0.025 * 44100 = 1102 samples, thus taking into account standard 1024(2^N) samples. In this experiment, the frame step for 15ms (around 661 samples) has provided good results so that the same frame overlap is maintained. The first 1024 sample frame starts at sample 0, the next 1024 sample frame starts at 661 and so on until it reaches the end of the speech file. To do so, pad it with zeros if the speech file is not broken into an even number of frames.

Eventually, selected in MFCC with the following parameters gives good accuracy than other 1)first order FIR high pass filter used for pre-emphasis system with filter coefficient ' a ' having value 0.95 2) hamming window with frame size 25ms and 60% overlapping 3) twenty cepstral coefficients.

5.1.2 Linear Predictive Coding (LPC) as a feature extraction method

In this research, the output of the LPC algorithm is verified by changing the order of the LPC algorithm. The remaining parameters are the same as those used in MFCC, such as frame size, pre-emphasis filter order and window type. The experimental result shows that the percent accuracy of speech recognition is a function of the order of the LPC analysis. The highest accuracy was obtained for order p=4. The following table 5.3 and the respective graph shown in figure 5.10 describe the effect of the predictor order on percentage accuracy.

Table 5.3: Effect of order of predictor on % accuracy

LPC	Recognition accuracy in %
For P=1	21.54
For P=3	27
For P=4	33.6
For P=5	30.45
For P=6	28.78

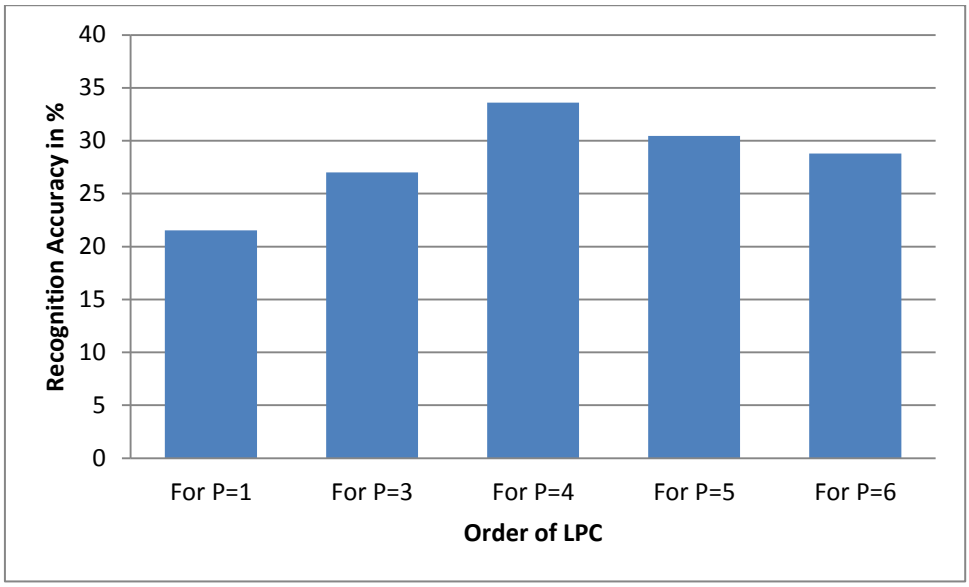


Figure 5.10: Effect of the LPC analysis on the accuracy of speech recognition

The significance of this method lies in both its ability to provide accurate estimates of speech parameters and its relative computational speed [66]. Analysis of the LPC is based on the assumption that the voice signal can be characterized by a predictor model that looks at past output values alone; thus it is an all-pool model in the Z transform domain [67].

The main difference between PLP and LPC analytical techniques is that the LP model assumes the vocal tract's all-pole transfer function with a specified number of resonances within the analytical band. The all-pole configuration of the LP approximates the distribution of power fairly well at all analysis band frequencies. This assumption is inconsistent with human hearing, because in the middle frequency range of the audible spectrum, the spectral resolution of hearing decreases with frequency and hearing is also more sensitive [68].

5.1.3 Relative Spectral Perceptual Linear Prediction (RASTA-PLP) as a feature extraction method

This study shows that RASTA PLP Cepstral analysis provides better result than RASTA PLP spectral analysis. Through ASR, the job is to interpret the language message through voice. This linguistic message is coded into vocal tract movements. These movements are reflected in the speech signal. The rate of change of non-linguistic speech components is often outside the typical rate of change of the shape of the vocal tract. The RASTA-PLP is taking advantage of this reality. It suppresses

spectral components that change slower or faster than the typical range of speech changes. In the presence of convolutionary and additive noise, RASTA-PLP processing increases the output of a recognizer [69]. The experimental results shown in Table 5.4 and Figure 5.11 show that the quality of the RASTA-PLP is better than the LPC analysis.

Table 5.4: Comparison of different PLP Features extraction techniques

Technique	Accuracy in %
RASTA PLP Spectral	39.45
RASTA PLP Cepstral	53.36

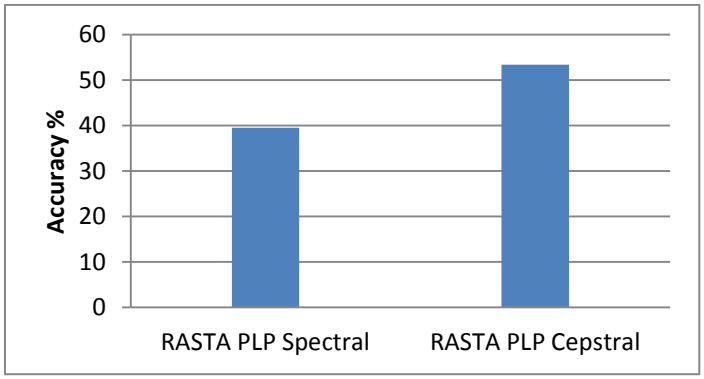


Figure 5.11: Comparison of different RASTA- PLP Features techniques

In first stage of experiment speech prediction was done without phoneme separation on different dataset, by deciding efficient feature extraction technique and Minimum Euclidean distance as classifier. Experiments were done on three different methods: Linear Predictive Coding (LPC), Relative spectral Perceptual Linear prediction (RASTA PLP) and Mel Frequency Cepstral Coefficients (MFCC). This research finds that MFCC feature extraction technique results better than LPC and RASTA-PLP since Mel-Frequency speech analysis is based on human perception studies. It only retains linguistic features, discards certain items that carry information like background noise, etc. [70].

As shown in Table 5.5 and Figure 5.12, the MFCC technique was used as the most appropriate technique for the extraction of features in the research work.

Table 5.5: Comparative analysis of feature extraction techniques

Sr. No.	Feature Extraction Technique	Recognition Accuracy %
1	Linear Predictive Coding (LPC)	35.25
2	Relative spectral Perceptual Linear prediction (RASTA PLP)	38.5
3	Relative cepstral Perceptual Linear prediction (RASTA PLP)	46.2
4	Mel Frequency Cepstral Coefficients (MFCC)	67

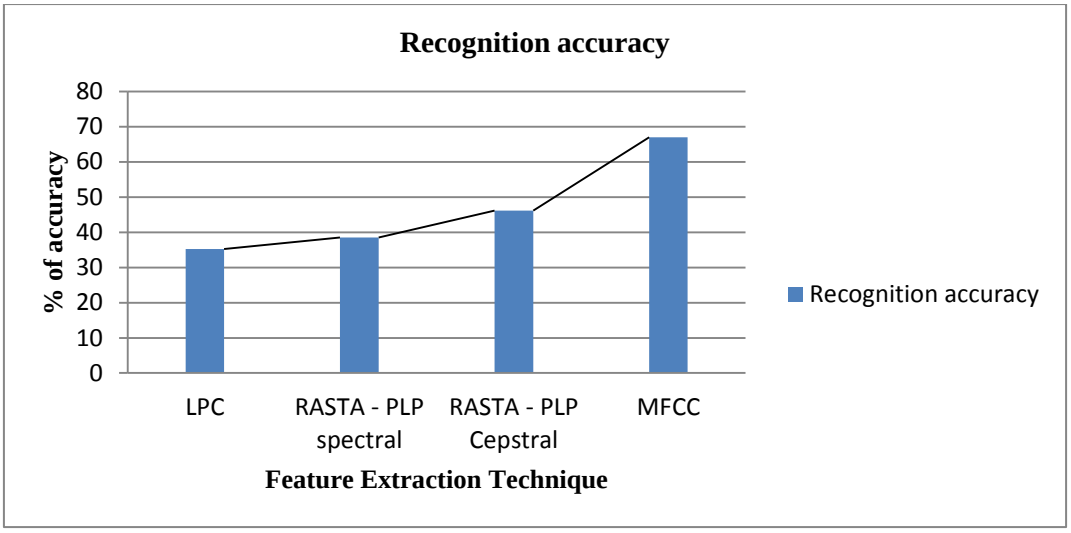


Figure 5.12: Comparison between Feature Extraction Techniques

As a result, the MFCC technique was used as the most appropriate technique for extraction of features in the research work. The same feature extraction technique with the same parameters was applied on various classifiers to determine the accuracy of speech recognition. For previous experiments, the right technique of feature extraction technique is used to find the correct word by using the minimum Euclidean distance

Classifier by using MFCC as a feature extraction tool, the following section discusses the performance results of different classifiers.

5.2 Performance of classifiers

The overall speech recognition system quality depends on the pre-processing techniques used, the chosen feature extraction technique and the classifiers used. Throughout the classification procedure, decisions are made using information relating to known patterns based on similarity measures from training patterns. They are then tested using the patterns that are unknown. Even though a speech recognition scenario involves a number of different classes, a multi-class classification technique is needed. The data were divided into training and test sets for virtually all classification methods. Every instance in the training set has a target value representing the correct category and a set of attributes. Test data do not include a target value. The goal of the classifier is to generate a training data model that predicts the target values of the test data [71].

Minimum Euclidean distance, Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), Hidden Markov Model (HMM) and Artificial Neural Network (ANN) classifiers are used in this research work.

5.2.1 Minimum Euclidean distance

The minimum Euclidean distance classifier is used to classify unknown speech data into classes that minimize the distance in multi-function space between the speech data and the class. The distance is known as a similarity index, so the minimum Euclidean distance is the same as the maximum similarity. Is used in situations where the disparity between the different classes is different. The distance between the Euclidean and the similarity index is technically equal.

The Minimum Euclidean distance algorithm is a very good algorithm for classifying static postures and recognition of non-temporal patterns. An especially it is fast classifier. The Minimum Euclidean distance algorithm's main limitation is that selecting the "wrong" number of clusters will result in poor classification performance. Therefore, the user may want to train the algorithm with several different cluster values to decide a

"good" cluster value. The table 5.5 and figure 5.12 shows the performance of Minimum Euclidean distance classifier applied on different feature extraction

techniques. The result shows that MFCC along with Minimum Euclidean distance gives good speech recognition accuracy than other feature extraction techniques.

5.2.2 Support Vector Machine (SVM) classifier

The goal of any learning machine is to achieve good performance in generalization, given a finite amount of training data, by striking a balance between the fitness goodness achieved on a given training dataset and the ability of the machine to achieve error-free recognition on other datasets. Based on this concept, support vector machines have been shown to achieve good performance in generalization without prior knowledge of the data.

An SVM's principle is to map the input data to a higher dimensional feature space that is not linearly related to the input space and determine a separate hyper plane with a maximum margin in the feature space between the two classes [77].

In sequential forward sequence (SFS) way, the features are selected. The experimental results show that the Gaussian medium SVM using 20 numbers of features provides the best accuracy compared to other features.

The predicted word accuracy (for different number of features selected) for all digits is calculated for different kernel function as shown in Table 5.6.and figure 5.13

Table 5.6: An Effect of kernels in SVM on average accuracy verses no of features for All Digits

Sr. No	Type of SVM	**Accuracy in %**						
		C=8	C=10	C=12	C=14	C=16	C=18	C=20
1	Linear SVM	50.09	51.09	51.27	52.09	53	52.45	53
2	Quadratic SVM	52.27	53.54	50.27	50.45	52.91	52.09	52.91
3	Cubic SVM	51.09	30.46	50.45	50.91	50.45	50.54	54.20
4	Fine Gaussian SVM	52.08	52.09	51.18	51.36	51.09	50.18	52.30
5	Medium Gaussian SVM	54.82	50	55.45	54.09	50.91	50.91	57.88
6	Coarse SVM	57.27	59.09	59.55	51.36	52	50.27	54.8

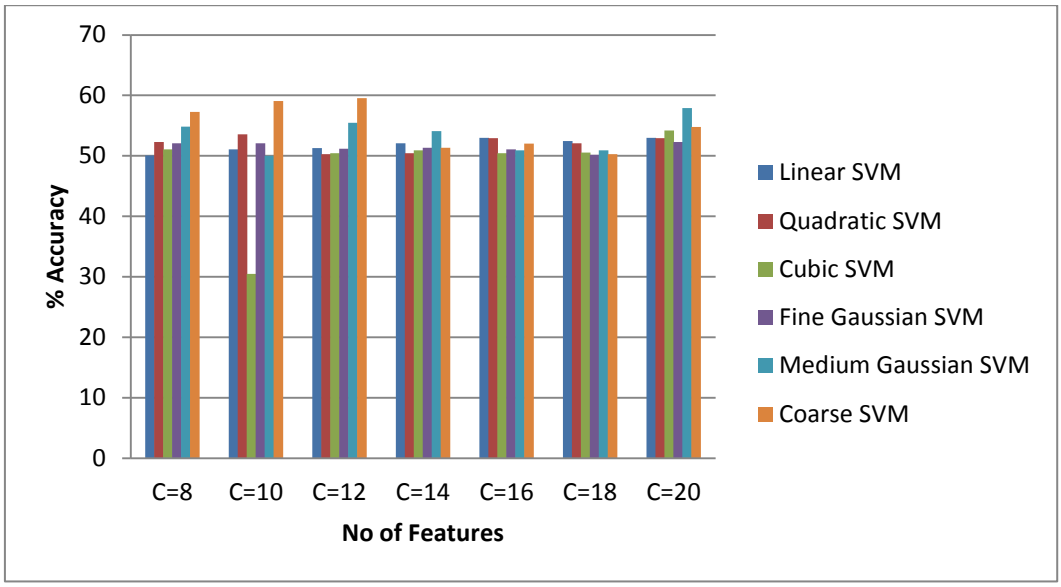


Figure 5.13: An Effect of kernels in SVM on average accuracy verses no of features for All Digits

For different types of kernels used in SVM, Table 5.7 and figure 5.14 shows the effect of validation on average accuracy for all digits. Using the 20 features, the experimental results show that the accuracy of the classifier is not affected by the validation process except for Quadratic SVM

Table 5.7: An Effect of validation in SVM on average accuracy for All Digits (No. of features C=20)

Sr. No	Type of SVM	Accuracy in %		
		No validation	5 fold cross validation	10 fold cross validation
1	Linear SVM	53	53	53
2	Quadratic SVM	52.91	52.20	52.20
3	Cubic SVM	54.20	54.20	54.20
4	Fine Gaussian SVM	52.30	52.30	52.30
5	Medium Gaussian SVM	57.88	57.88	57.88
6	Coarse SVM	54.8	54.8	54.8

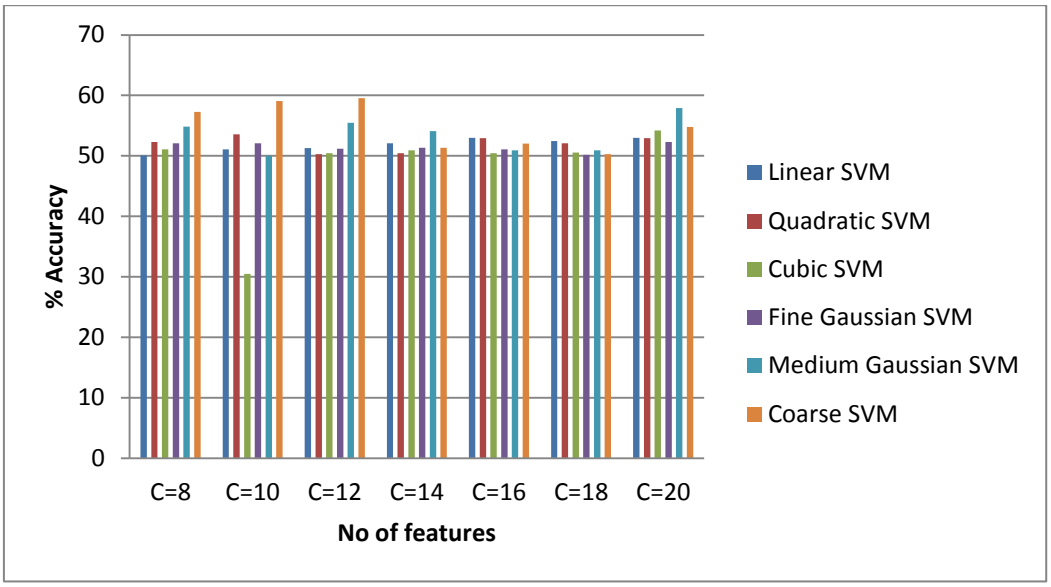
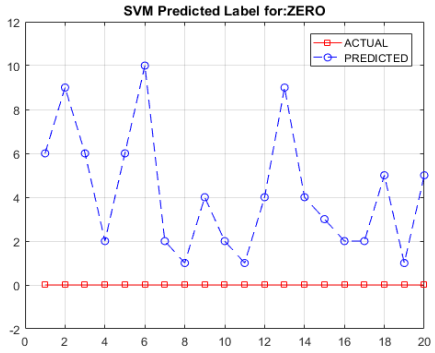
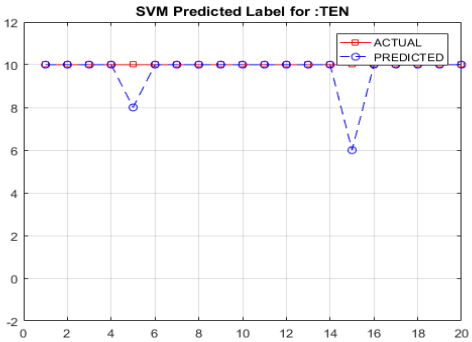


Figure 5.14: An Effect of validation in SVM on average accuracy for All Digits (No. of features C=20)

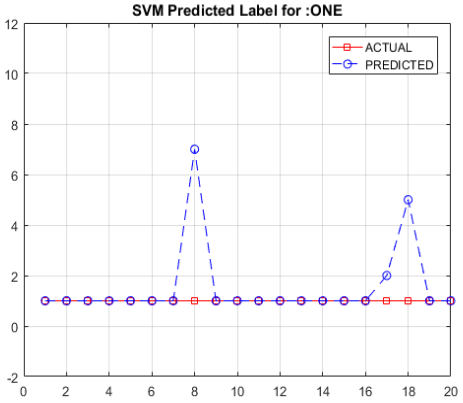
The following graph 5.15 (a)-(k) shows the relationship between actual and predicted word without phoneme separation using SVM classifier for digits zero to ten.



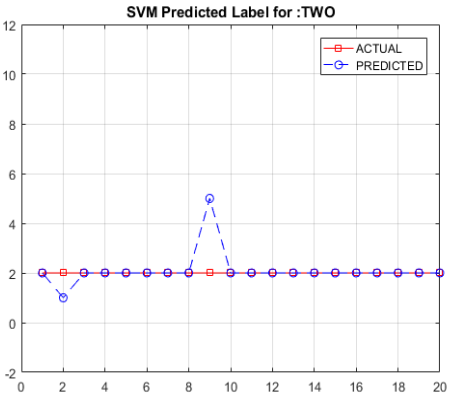
(a)



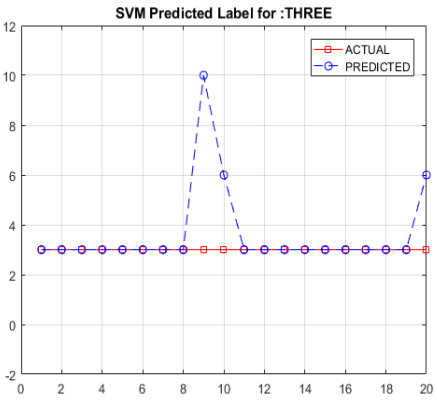
(b)



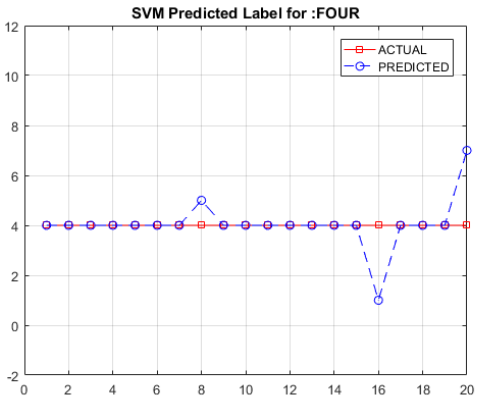
(c)



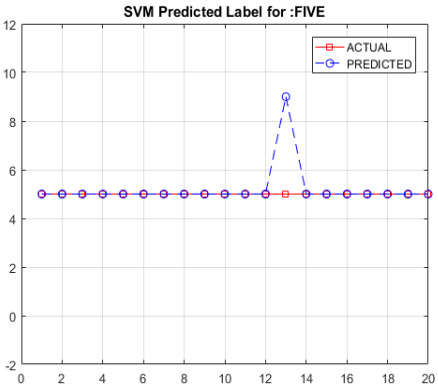
(d)


SVM Predicted Label for :THREE
ACTUAL
PREDICTED

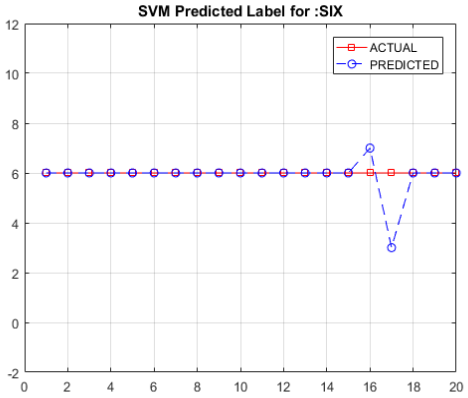
(e)


SVM Predicted Label for :FOUR
ACTUAL
PREDICTED

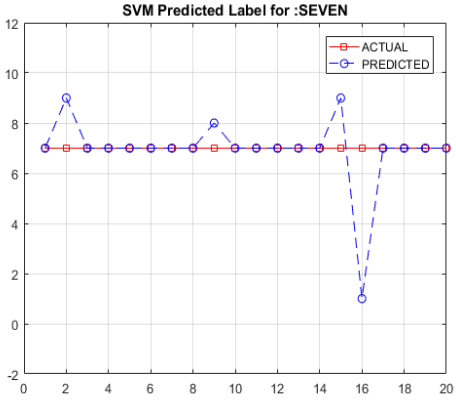
(f)


SVM Predicted Label for :FIVE
ACTUAL
PREDICTED

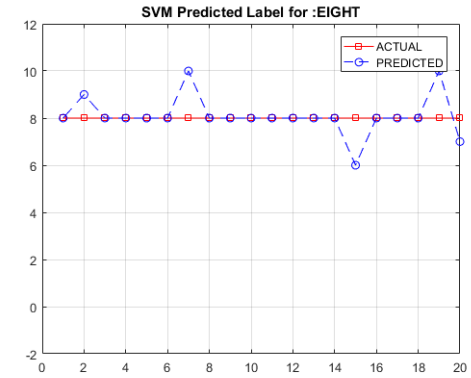
(g)


SVM Predicted Label for :SIX
ACTUAL
PREDICTED

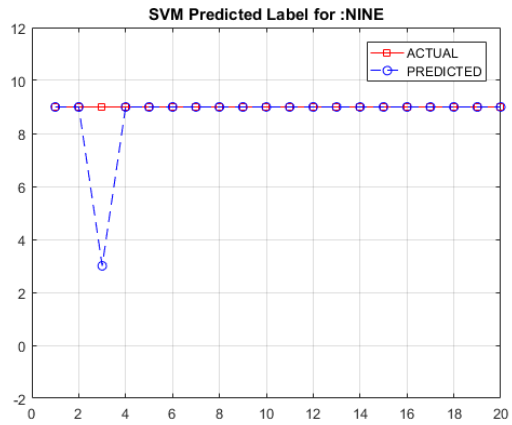
(h)


SVM Predicted Label for :SEVEN
ACTUAL
PREDICTED

(i)


SVM Predicted Label for :EIGHT
ACTUAL
PREDICTED

(j)



(k)

Figure 5.15 (a)-(k): Relationship between actual and predicted word without phoneme separation using SVM classifier

Above result shows that SVM classifier is not suitable classifier to recognize this particular data.

5.2.3 k-Nearest Neighbor (k-NN) classifier

One of the instance-based methods is the K-Nearest Neighbor classifier, which is also called a lazy algorithm. Statistical classifiers are based on the decision rule of Bayes and can be divided into parametric and non-parametric classifiers [75].

The time required to check a pattern in K-NN depends on the number of training samples m and the size of the O (n*m) feature vector n [76].

In this research, the system is tested for various values of k with different distance measurement. Table 5.8 and figure 5.16 Shows that the value of k increases, the predicted word's average accuracy decreases.

Table 5.8: Effect of k-value on average accuracy for All Digits (No. of features=20)

Sr. No.	Distance Measures	% of Average Accuracy		
		k=1	k=3	k=3
1	Euclidean	73.00	73.00	73.00
2	Jaccard	12.73	10.45	9.55
3	Cityblock	62.18	61.06	61.04
4	Seuclidean	62.18	62.27	61.82
5	Hamming	12.73	10.45	9.55
6	Chebychev	60	57.72	55.45
7	Cosine	62	62.09	61.03
8	Mahalanobis	61.82	62.27	60.45
9	Minkowski	60.18	62.27	61.82
10	correlation	61.82	60.73	60.45

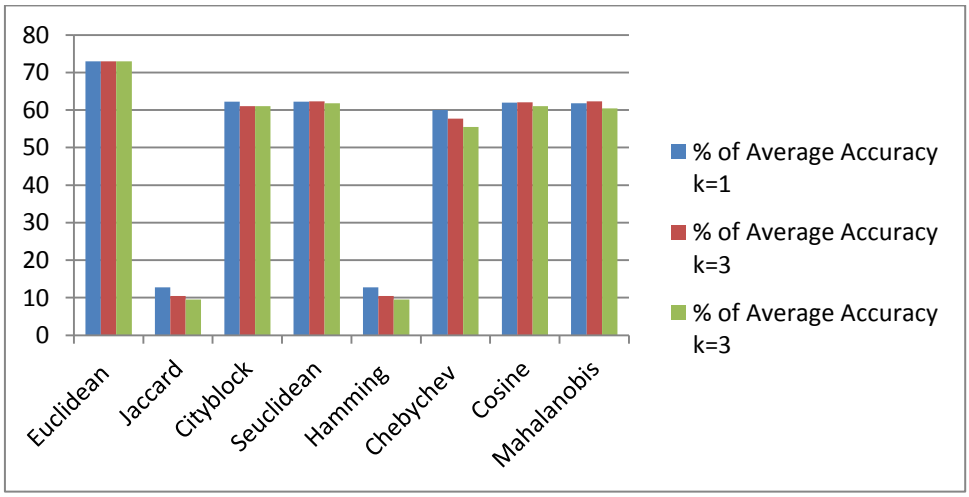


Figure 5.16: Effect of k-value on average accuracy for All Digits (No. of features=20)

Euclidean distance measurement for all k values provides better average accuracy than other distance measurements. However, the distance measurements of Hamming and Jaccard perform the worst for the discourse disorder database.

For different types of k-NN, Table 5.9 and figure 5.17 shows the effect of validation on average accuracy for all digits. The 20 characteristics are used.

Table 5.9: Effect of validation in k-NN on average accuracy for All Digits

(No. of features=20 and Euclidean distance measure)

Sr. No	Type of KNN	Accuracy in %		
		No validation	5 fold cross validation	10 fold cross validation
1	Fine KNN	73	73	73
2	Medium KNN	60.45	60.45	60.45
3	Coarse KNN	54.09	54.09	54.09
4	Cosine KNN	66.82	66.82	66.82
5	Cubic KNN	66.36	66.36	66.36
6	Weighted KNN	65.91	65.91	65.91

The experimental results show that the validation process does not affect the classifier's accuracy. Fine, weighted k-NN provides better precision. The experimental results show that the validation process does not affect the classifier's accuracy. Fine, weighted k-NN provides better precision.

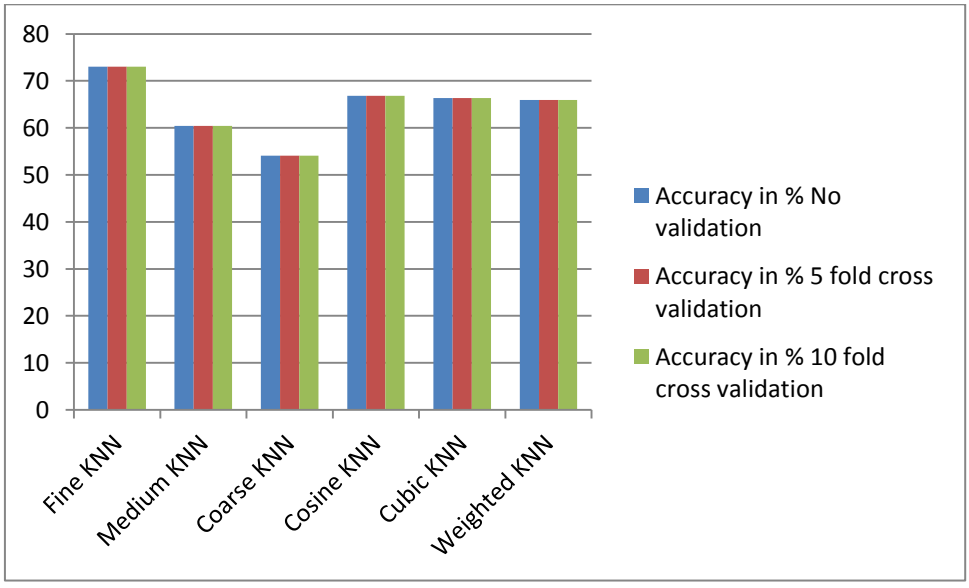
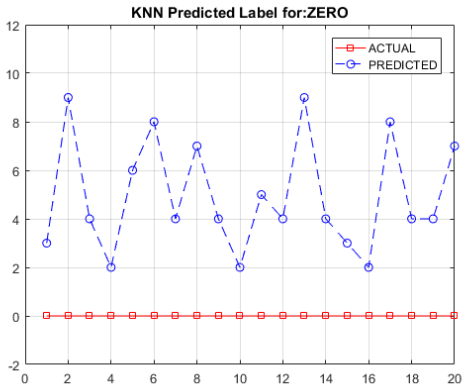
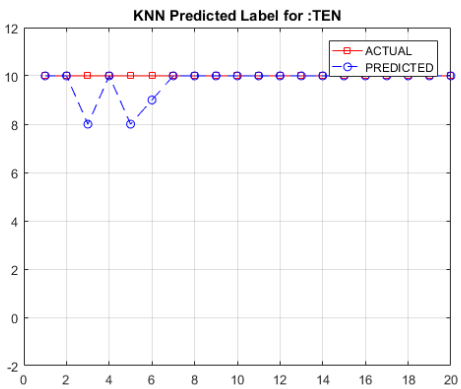


Figure 5.17: Effect of k-value on average accuracy for All Digits (No. of features=20)

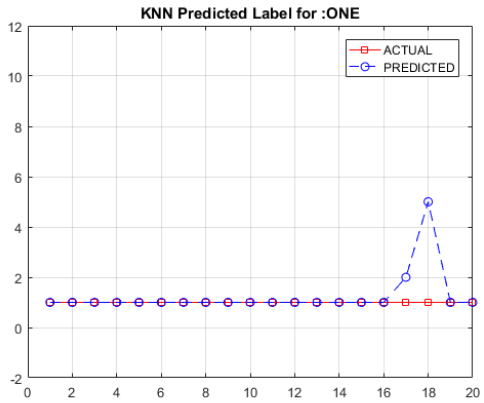
The following figure 5.18 (a to k) shows the relationship between actual and predicted word without phoneme separation using KNN classifier for digits zero to ten.



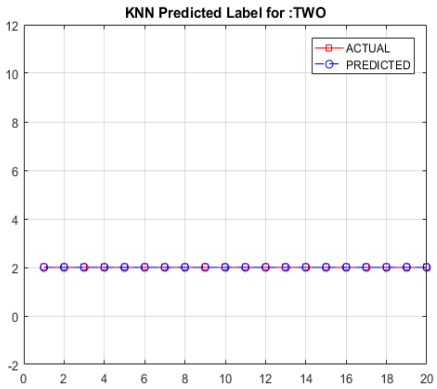
(a)



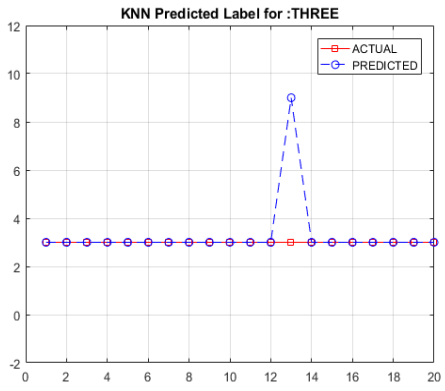
(b)



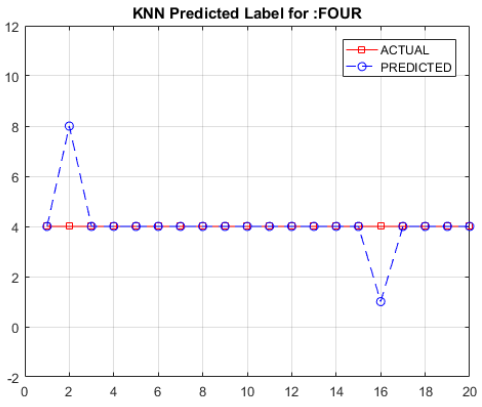
(c)



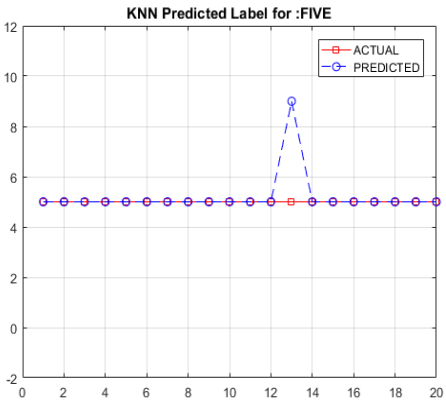
(d)



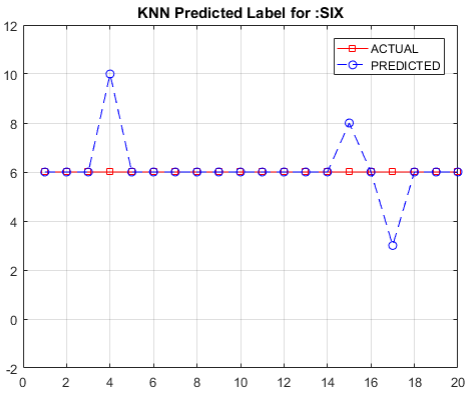
(e)



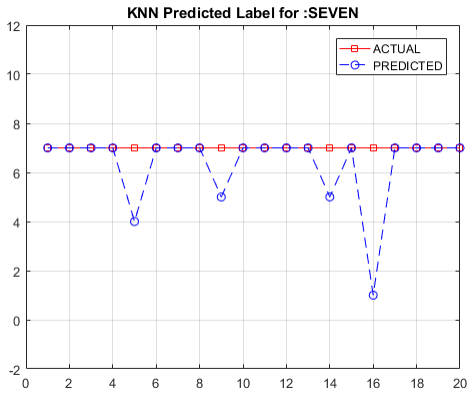
(f)



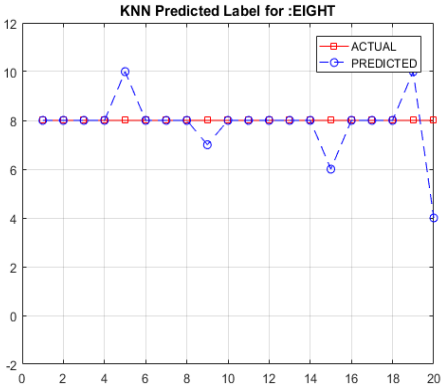
(g)



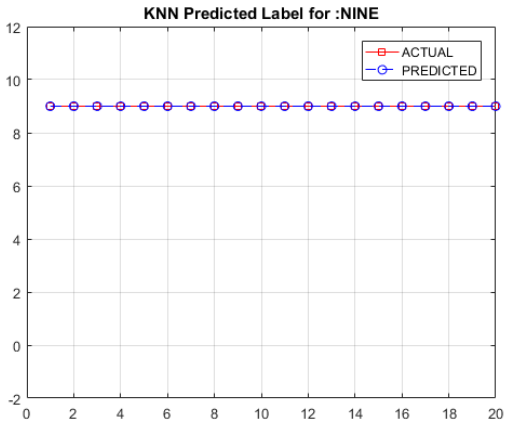
(h)



(i)



(j)



(k)

Figure 5.18 (a)-(k): Relationship between actual and predicted word without phoneme separation using KNN classifier

Above result shows that most of the actual words are predicted correctly.

Table 5.10 shows that overall performance of k-NN classifier is better than SVM for all digits.

Table 5.10: Overall performance of classifier on average accuracy for All Digits

Sr. No.	Classifier	No. of features used	Accuracy in %
1	Fine KNN	20	73
2	Medium Gaussian SVM	20	57.88

K-NN (Kernel method) is better than SVM (statistical method) for the sake of certain facts mentioned below.

In case of SVMs, the parameters of one class are trained on the samples of all classes. But for k-NN statistical classifiers, the parameters of one class are estimated from the samples of its own class only. Such two classifiers contrasted in the following respects on the basis of the characteristics.

- Complexity of learning activities:

K-NN classifier parameters are typically modified by measuring distance. The training time is proportional with the number of samples by feeding the training samples a fixed number of sweeps. Quadratic programming (QP) learning is conducted on SVMs, and training time is usually equal to the square number of samples.

- Training's flexibility:

In feature weighting, the parameters of K-NN classifiers can be adjusted to improve classification accuracy for global performance and an existing classifier can also be easily added to a new class. On the other side, SVMs can only be conditioned on the basis of systemic trends. The classifier is proportional to the number of classes and includes re-training with all samples to ensure the accuracy of parameters, adding new classes or new samples.

- Precise classification:

In many studies, superior classification accuracies were demonstrated to KNN classifiers by SVMs. If learning with enough specimens, the SVM classifiers give higher accuracy than the numerical classifiers.

- Space of learning difficulty and complexity of execution:

SVM learning by Quadratic Programming also results in a large number of SVs at the same rate of classification accuracy, which should be processed and measured in classification. K-NN classifiers have far fewer parameters and can be easily controlled. In a word, less space and computation consumes K-NN classifiers than SVMs.

5.2.4 Hidden Markov Model (HMM) classifier

Many speech recognition systems, like the Hidden Markov Model, use probability-based model. HMM is effective in modeling variations in speaking time and frequency domain statistical parameters. HMM also has a major advantage in rapidly estimating the parameters from training data. HMM is a community of transition-related states. It starts with the initial state. In the finite time, the state moves from one state to another, known as transition state, and in that state, one output sequence is generated. The sequences of output and transition state are randomly related, which is determined by distribution of probability. In the case of HMM, the series of visible states produced over time can be observed by a term and the transition of the visible state sequence over time is hidden.

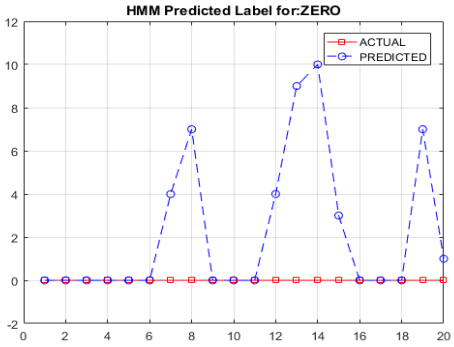
The recognition of average accuracy for overall digits found in evaluation for disordered speech is given in Table 5.11.

TABLE 5.11: % OF AVERAGE ACCURACY USING HMM

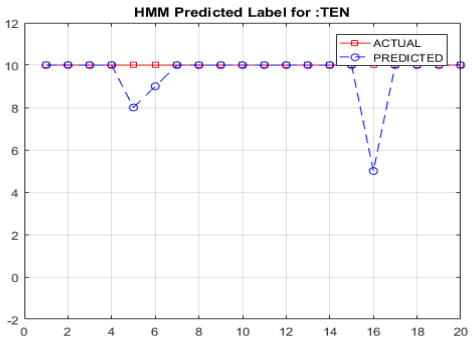
No. of Features used	8	12	16	20
% of average accuracy	45.76	53.4	56.8	58.9

HMM has one downside that every speech frame is assumed to be independent of its neighbors. Nevertheless, it is not literally a reality. It therefore finds less accuracy than other classifiers.

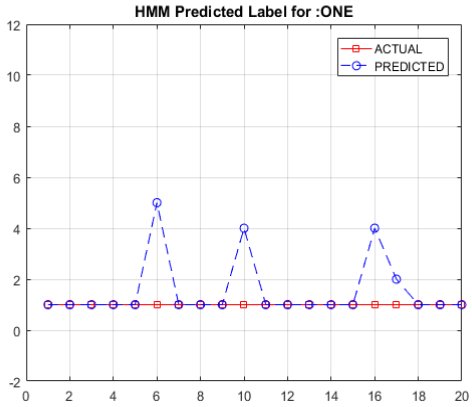
The following figure 5.19 (a to k) shows the relationship between actual and predicted word without phoneme separation using HMM classifier for digits zero to ten.



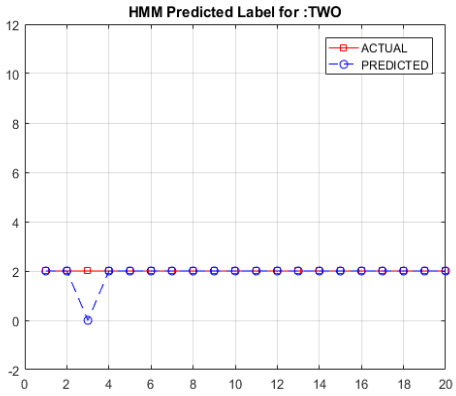
(a)



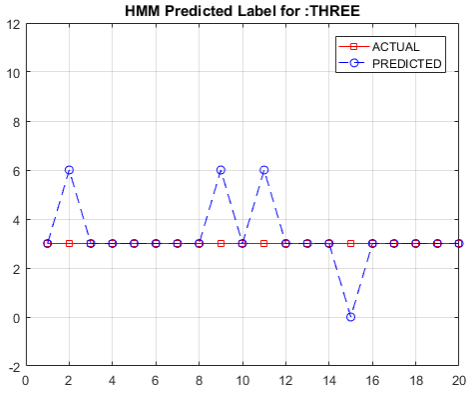
(b)



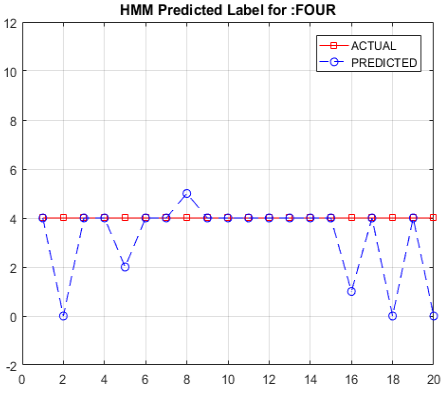
(c)



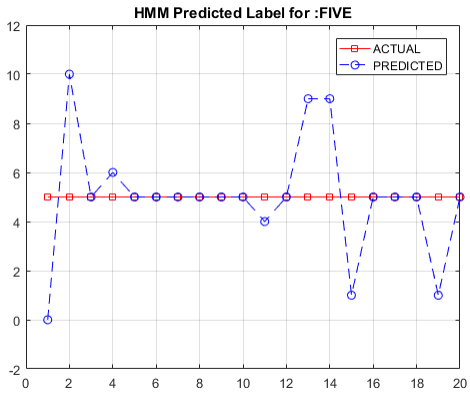
(d)



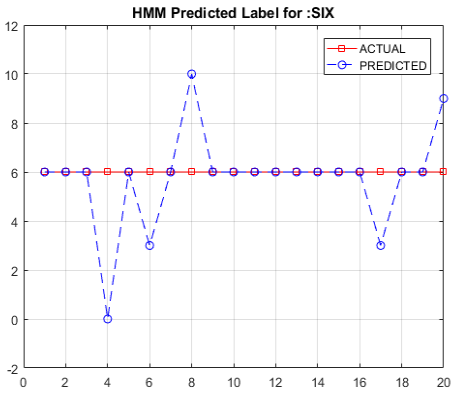
(e)



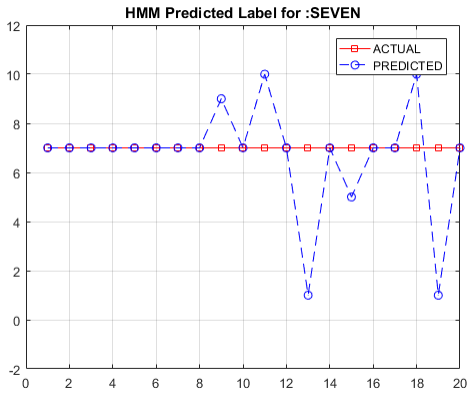
(f)



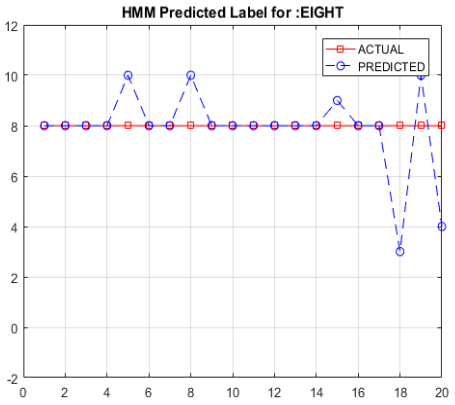
(g)



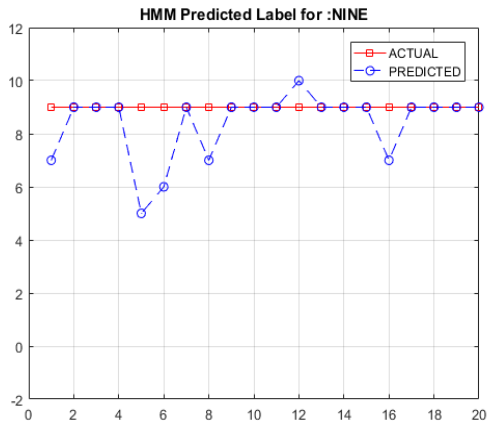
(h)



(i)



(j)



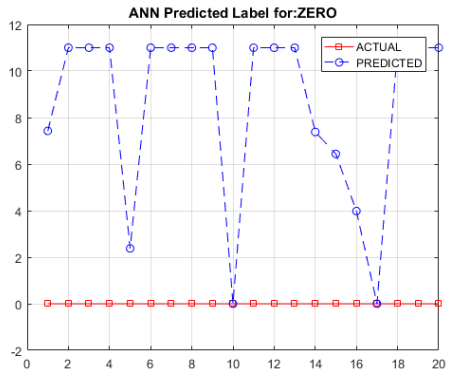
(k)

Figure 5.19 (a)-(k): Relationship between actual and predicted word without phoneme separation using HMM classifier

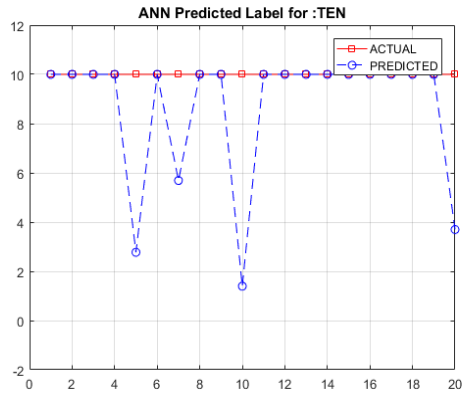
5.2.5 Artificial Neural Network (ANN) classifier

Due to their parallel distributed storage, distributed memory, error consistency, and ability to learn and differentiate patterns, ANNs are used in many applications. ANN is a system for processing information which consists of a number of basic units or nodes called neurons. Each neuron accepts a weighted array of inputs and produces an output [78]. ANN-based algorithms are well adapted to tackle tasks of speech recognition. Neural network models, inspired by the human brain, use a variety of features such as learning, generalization, adaptive, fault tolerance, etc.[79]. The architecture of the MLP network, consisting of an input layer, one or more hidden layers, and an output layer, has been used in this research work. The algorithm used is the learning algorithm for back propagation. The input is provided to the network in this type of network and passes to the output layer through the weights and nonlinear activation functions, and the error is corrected in a backward direction using the well-known error propagation correction algorithm.

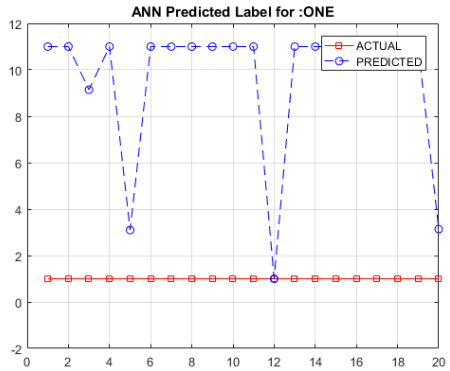
The following figure 5.20 (a)-(k) shows the relationship between actual and predicted word without phoneme separation using ANN classifier for digits zero to ten.



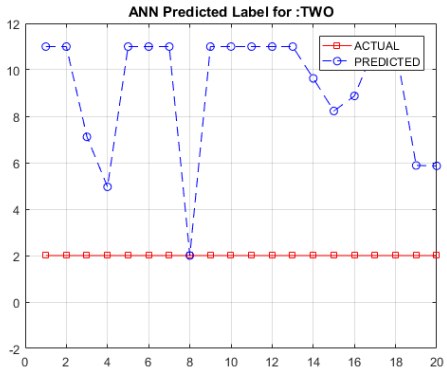
(a)



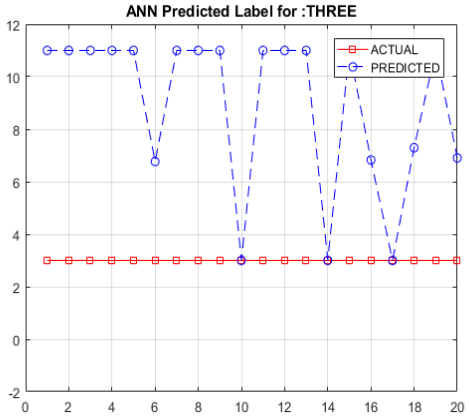
(b)



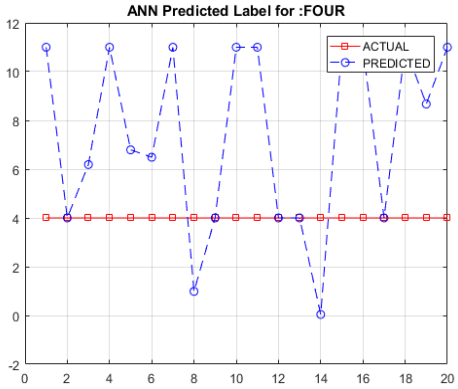
(c)



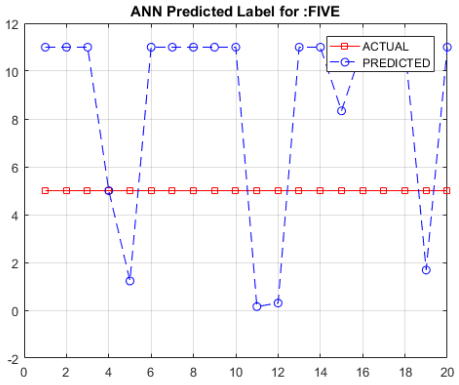
(d)


ANN Predicted Label for :THREE
ACTUAL
PREDICTED

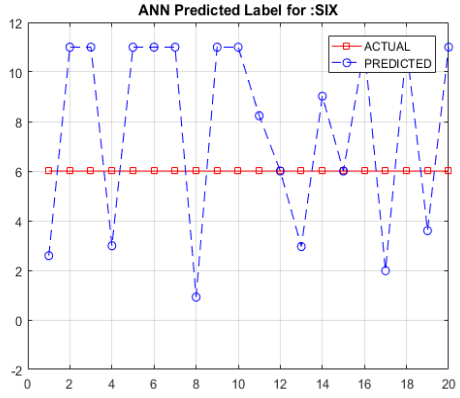
(e)


ANN Predicted Label for :FOUR
ACTUAL
PREDICTED

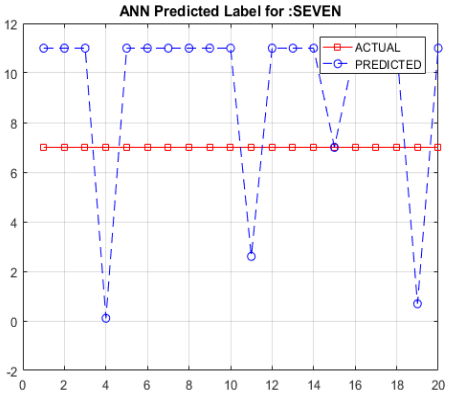
(f)


ANN Predicted Label for :FIVE
ACTUAL
PREDICTED

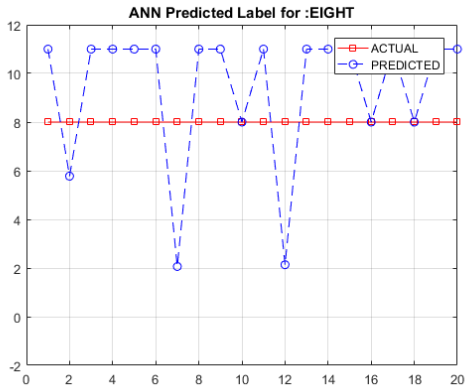
(g)


ANN Predicted Label for :SIX
ACTUAL
PREDICTED

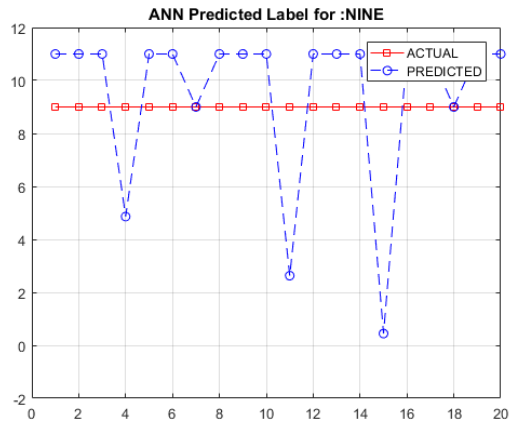
(h)


ANN Predicted Label for :SEVEN
ACTUAL
PREDICTED

(i)


ANN Predicted Label for :EIGHT
ACTUAL
PREDICTED

(j)



(k)

Figure 5.20 (a)-(k): Relationship between actual and predicted word without phoneme separation using ANN classifier

The result shows that the accuracy of recognition is lower than that of other classifiers as it has few limitations, such as [80]

1) There is no formal approach in ANN

2) ANN can provide an unpredictable value of output

3) There is no explanation of the problem solving approach of many ANN systems.

4) A black box of nature.

5) Empirical character for the development of model.

6) There is more time required for training the network for vast data and inability to accommodate the time sequence of speech.

5.2.6 Overall Performance of Classifiers

The method performance is then tested with MFCC as a feature extraction technique and using various classifiers such as the Minimum Euclidean Distance Classifier, k-NN, SVM, HMM and ANN. The findings of the test were summarized in Table 5.12 and figure 5.21.

Table 5.12: Comparative analysis of classifiers

Sr. No.	Type of Classifier	Recognition Accuracy %
1	Minimum Euclidean Distance Classifier	67
2	Artificial Neural Network	55
3	Hidden Markov Model	51
4	k Nearest Neighbor	73
5	Support vector machine	57.88

Results show k-NN classifier gives better accuracy as compared to other techniques.

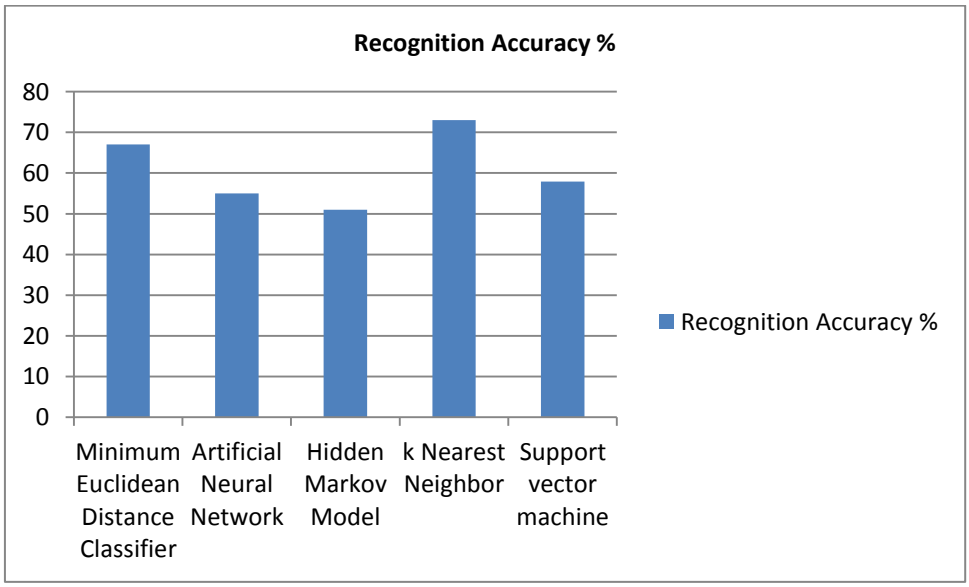


Figure 5.21: Comparison of k-NN with Other Classifier

The experimental result seen in Figure 5.21 and Table 5.12 shows that the k-NN classifier performs well for this particular system as compared to the other classifier.

It is therefore decided to use these two techniques to test the performance of the proposed system.

5.3 With phoneme separation

- The speech prediction algorithm has been developed using the separation of the phoneme.
- The developed algorithm was applied using MFCC as a feature extraction technique and k-NN classifier.
- The developed algorithm was used to separate the input digits into 2, 3, 4 and 5 segments.
- Conducting the test on input data showed improvement in accuracy, sensitivity, specificity and F-score with reduction in error rate.
- The next phase of experimental analysis is prediction of speech with separation of phonemes using segmentation on various data-sets. At this stage, input speech is divided into 2, 3, 4 and 5 segments for the identification of phonemes. Upon segmentation, the features are extracted using MFCC while retaining the same parameters used in previous experiments. Features extracted using MFCC are given as input to the k-NN classifier, which is used to identify the phoneme class. Correlation and occurrence of phonemes were observed using a positive position search algorithm. Using text to speech conversion technique, the word correctly predicted is converted into voice.

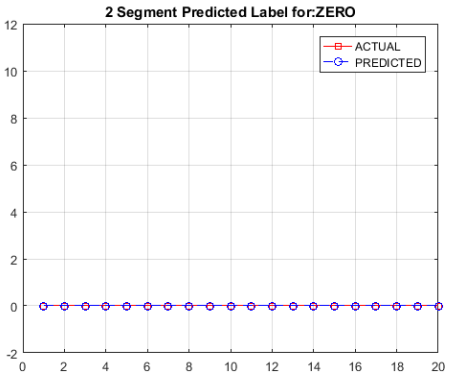
5.3.1 Results using 2 Segment

According to the lexicon table, one to ten digits shall consist of at least two and a maximum of five phonemes. Unknown speech is segmented into 2, 3,4 and 5 segments. The phonemes of each word are equally spaced. The segment of each word (phoneme) is compared to the data stored during the training of the respective segment type (such as 2 test segments with two training segments and so on). After that, by using a positive position search algorithm, find the correlation and occurrence of each phoneme. Maximum correlation and proper occurrence of phoneme is identified as a correct phoneme. Table 5.13 shows that how unknown speech of articulatory handicapped people is segmented into number of 2,3,4 and 5 segments .

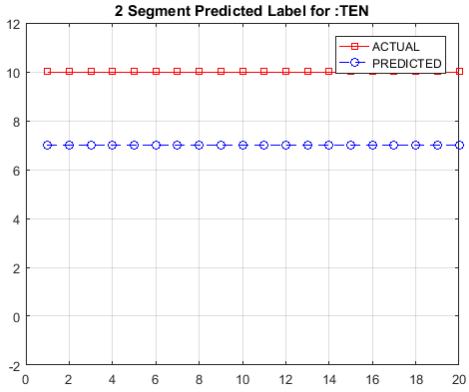
Table.5.13: Phoneme equivalence

Digi	Standard	Two	Three	Four segment	Five segment
0	Z IY R OW	IY R	IY F OW	N IY OW R	S K R OW Z
0	Z IY R OW	Z R	IY W V	Z IY R N	TH IY R OW N
0	Z IY R OW	Z R	TH R W	Z R R V	Z TH R R W
1	W AX N	OW AX	N N N	F AY EH N	T OW AY N S
1	W AX N	W AX	T TH TH	OW R IY EH	S OW AX IY Z
1	W AX N	T AX	T V S	T AX EH N	Z OW AX N OW
2	T UW	T AY	T UW N	T R V Z	S V R UW Z
2	T UW	N W	T UW AY	S V OW S	S F OW V S
2	T UW	T R	UW T N	F IY R N	W R R UW N
3	TH R IY	TH EH	TH TH IH	S T R S	S IH AY R S
3	TH R IY	TH R	IH TH IY	TH R IH IY	TH R R AY S
3	TH R IY	TH R	N N IY	T R R IY	IY R R R IY
4	F OW R	F OW	F F R	R OW AX N	S OW OW V R
4	F OW R	S OW	F OW N	S UW AX N	S F OW AX N
4	F OW R	F OW	R UW N	F OW AX N	F OW OW T N
5	F AY V	F AY	F V R	F AY AY R	S AY R UW R
5	F AY V	F AY	F EY Z	S AY IY IY	F AY AY AY N
5	F AY V	F V	F N T	F EH IY R	F EH V R S
6	S IH K S	F IH	IH S S	S IH K S	S K IH IH S
6	S IH K S	S IH	TH S N	S IH K S	S IH K V S
6	S IH K S	S K	S N S	S IH K N	S IH EH S N
7	S EH V EH N	EH V	EH F Z	S V AY IY	S EH V EH N
7	S EH V EH N	T AY	IH N F	S V EH N	T V EH EH S
7	S EH V EH N	S R	S UW V	S AY V N	T EH V EH N
8	EY T	S EH	EY T N	T EH EH W	S EH EH T S
8	EY T	EY EH	EY N IY	W EH T N	S EH EH N OW
8	EY T	Z EH	EY T V	S K S IH	S N V T N
9	N AY N	N AY	N EH EY	N AY R S	T AX AY AY TH
9	N AY N	N AY	IH TH N	N AY IH IY	S AY AY N S
9	N AY N	N AY	R N T	N AY EH N	F AY AY N S
10	T EH N	T EH	TH S T	S EH IH N	R IY EH EH S
10	T EH N	T EH	T EY N	OW EH K Z	T T EH AX TH
10	T EH N	T EH	R N N	TH EH EH OW	TH EY EH N S

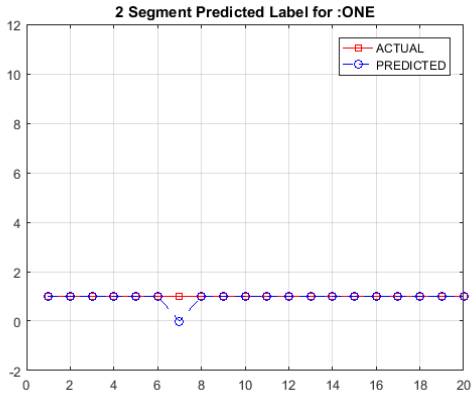
The following figure 5.22 (a to k) shows the relationship between actual and predicted word with 2 segment phoneme separation using KNN classifier for digits zero to ten.



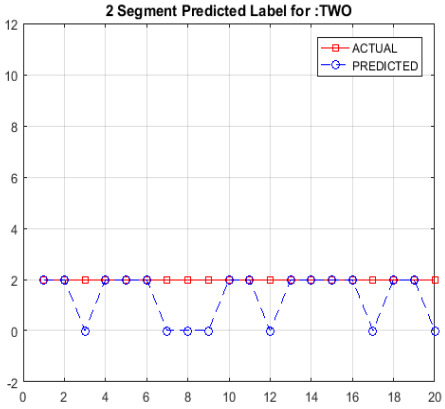
(a)



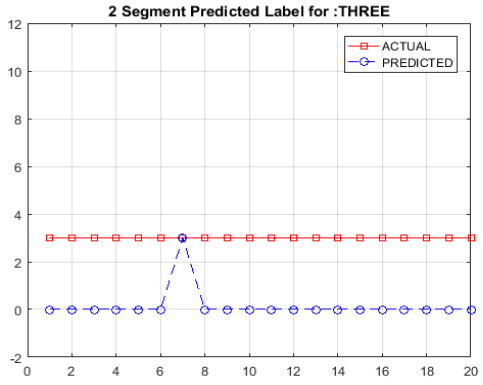
(b)



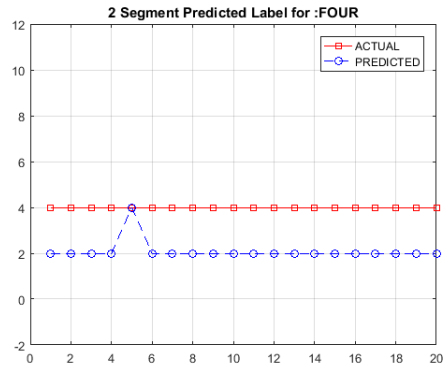
(c)



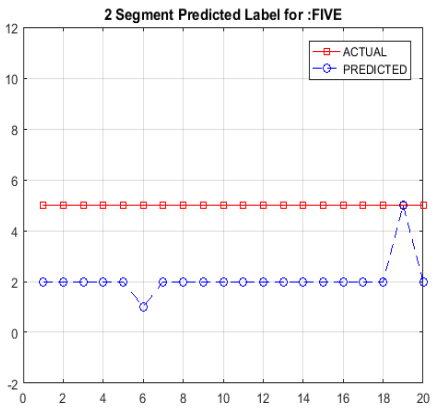
(d)

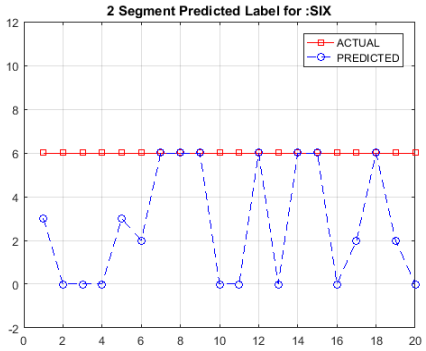


(e)



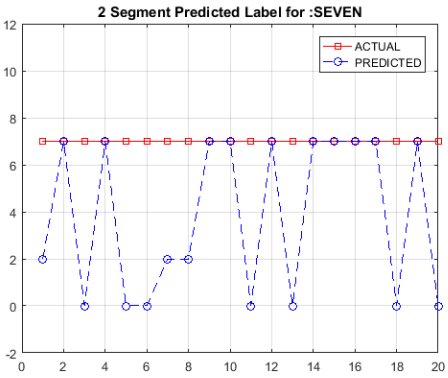
(f)

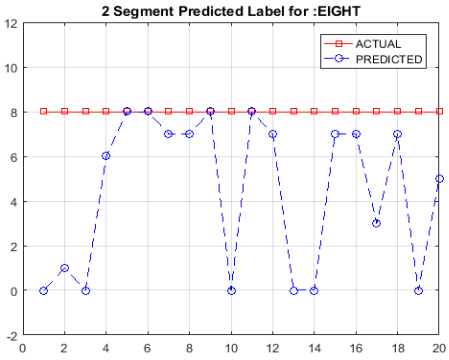




(g)

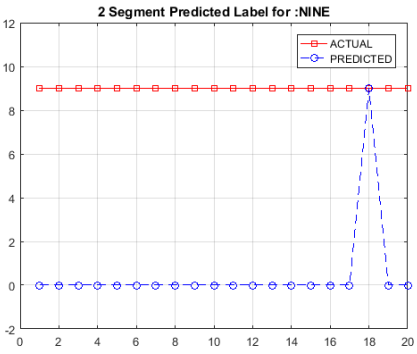
(h)





(i)

(j)



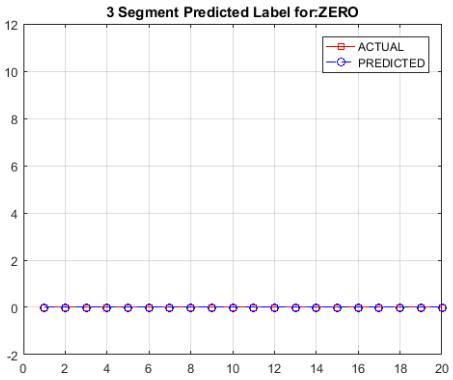
(k)

Figure 5.22: (a)-(k): Relationship between actual and predicted word with 2 segment phoneme separation using KNN

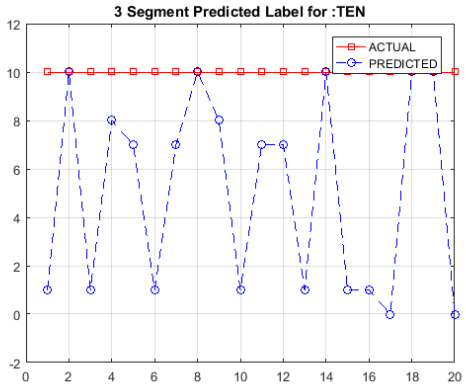
The result shows that the separation of two segments of the phoneme does not produce accurate results.

5.3.2 Results using 3 Segment

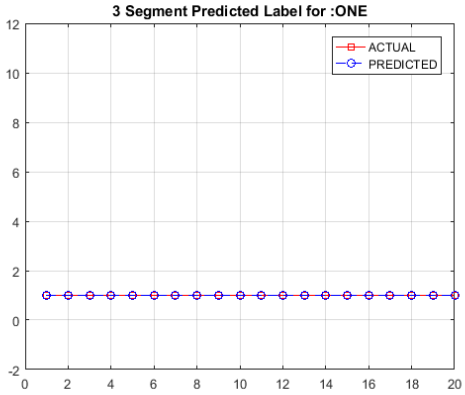
The following figure 5.23 (a to k) shows the relationship between actual and predicted word with 3 segment phoneme separation using KNN classifier for digits zero to ten.



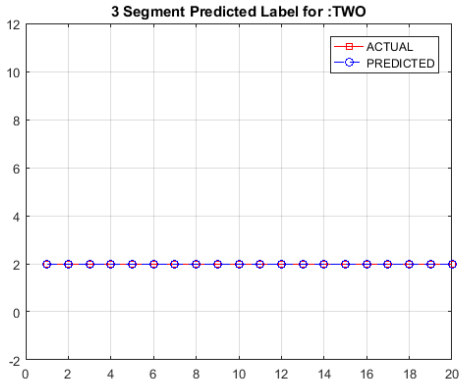
(a)



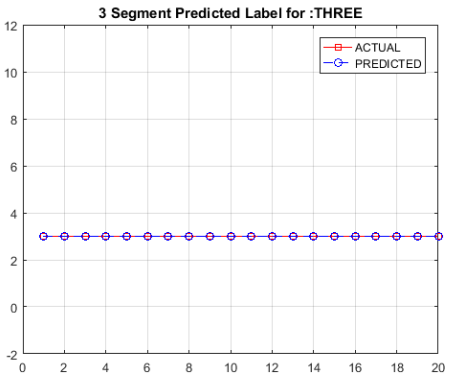
(b)



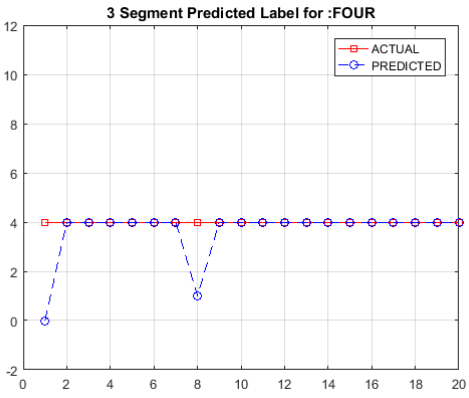
(c)



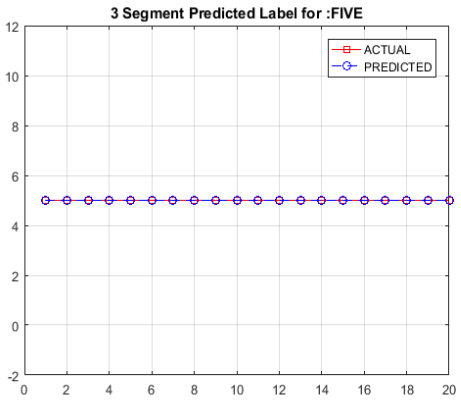
(d)



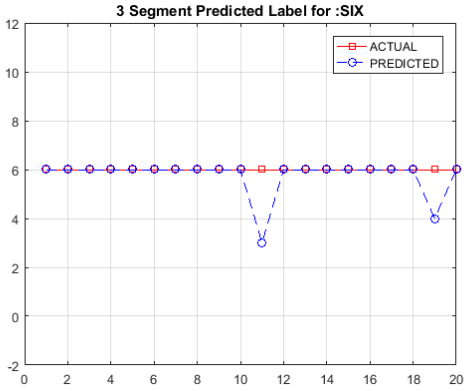
(e)



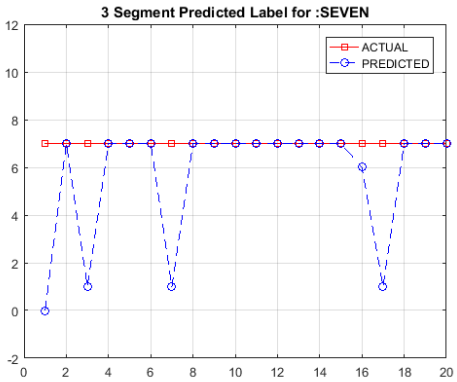
(f)



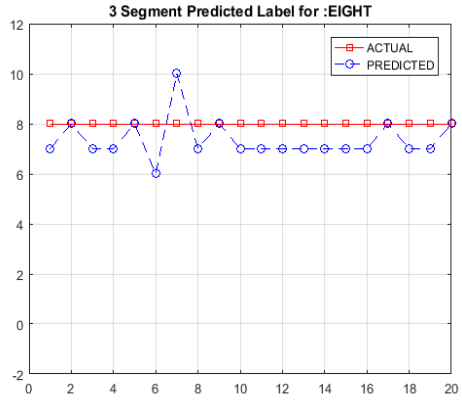
(g)



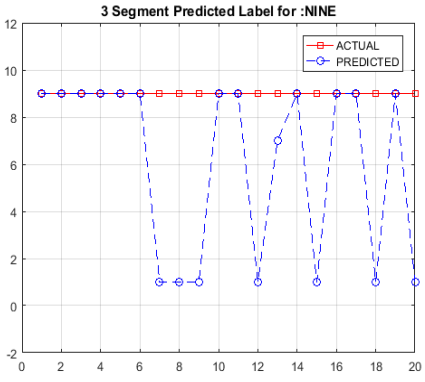
(h)



(i)



(j)

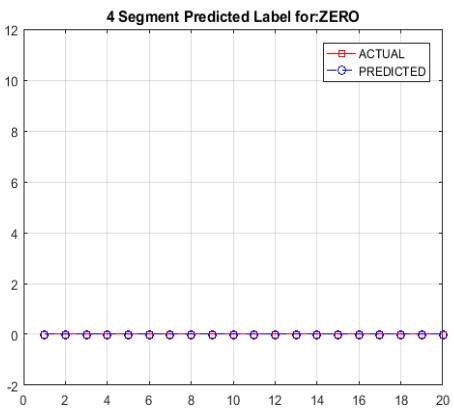


(i)

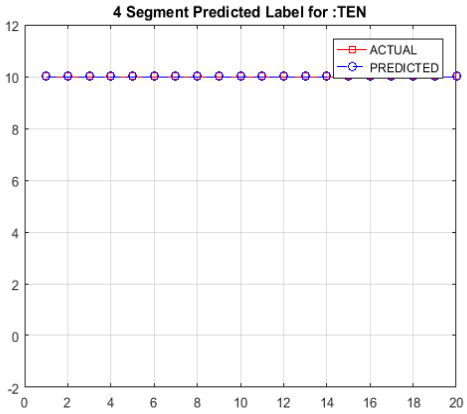
Figure 5.23 (a)-(k): Relationship between actual and predicted word with 3 segment phoneme separation using KNN

5.3.3 Results using 4 Segment

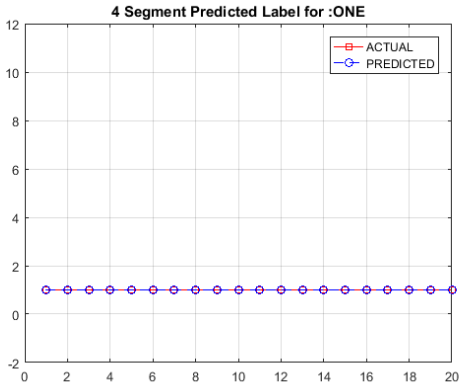
The following figure 5.24 (a to k) shows the relationship between actual and predicted word with 4 segment phoneme separation using KNN classifier for digits zero to ten.



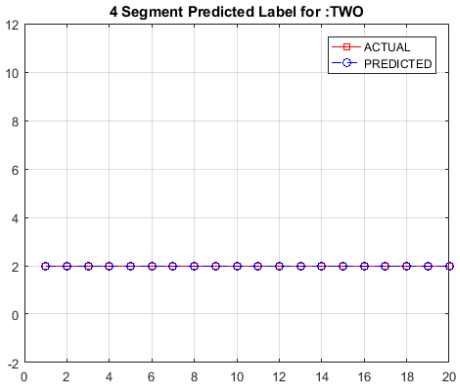
(a)



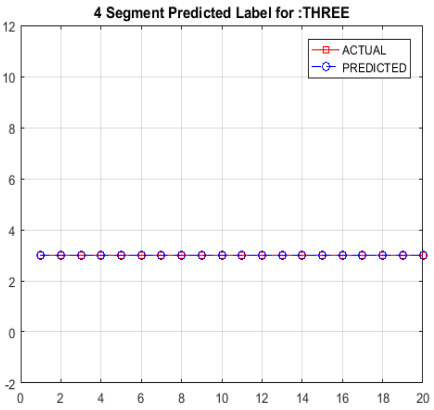
(b)



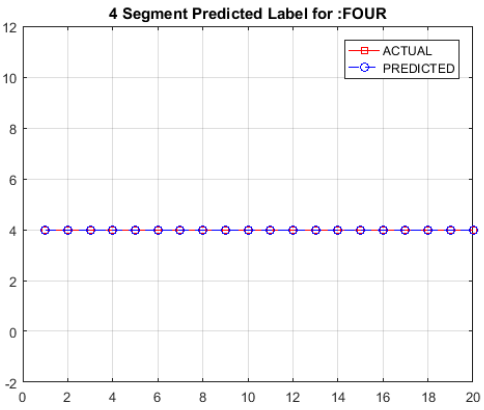
(c)



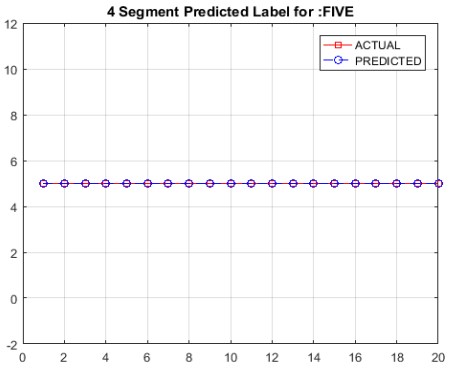
(d)



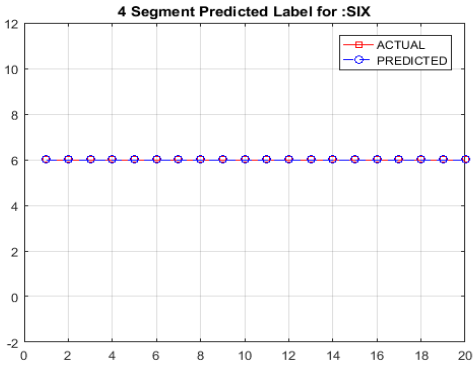
(e)



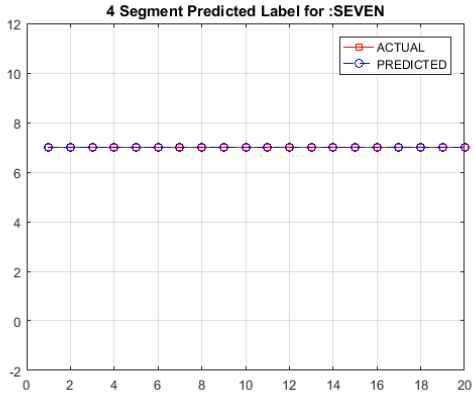
(f)



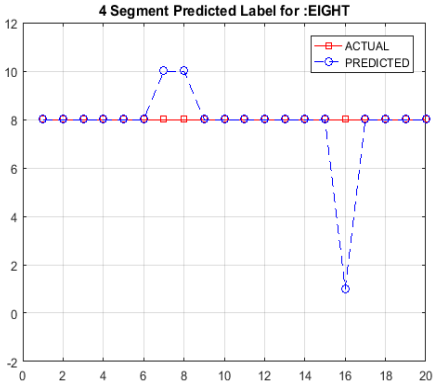
(g)



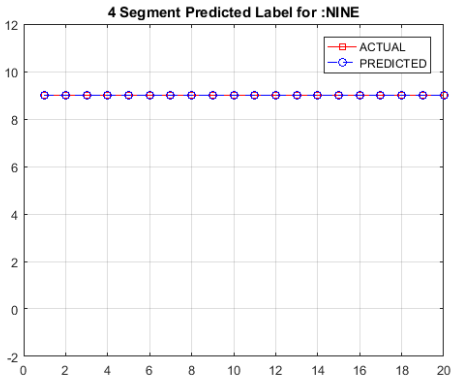
(h)



(i)



(j)

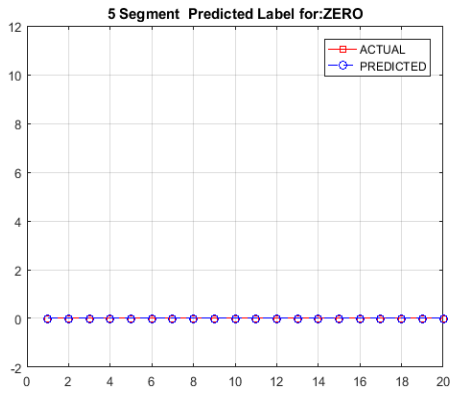


(k)

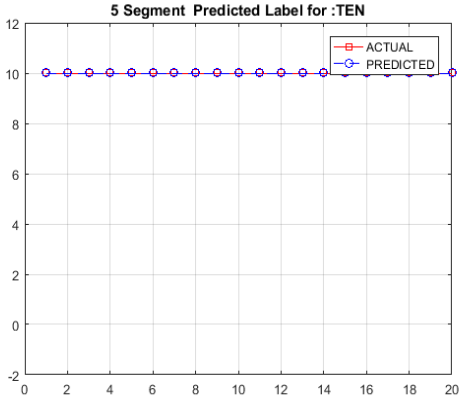
Figure 5.24 (a)-(k): Relationship between actual and predicted word with 4 segment phoneme separation using KNN

5.3.4 Results using 5 Segment

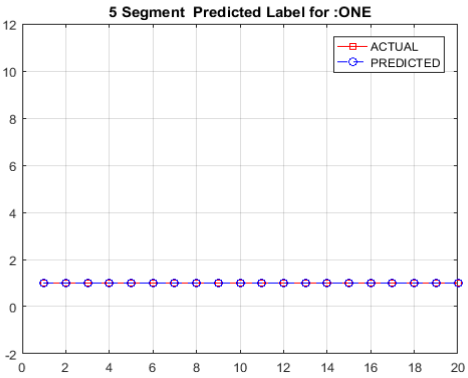
The following figure 5.25 (a to k) shows the relationship between actual and predicted word with 5 segment phoneme separation using KNN classifier for digits zero to ten.



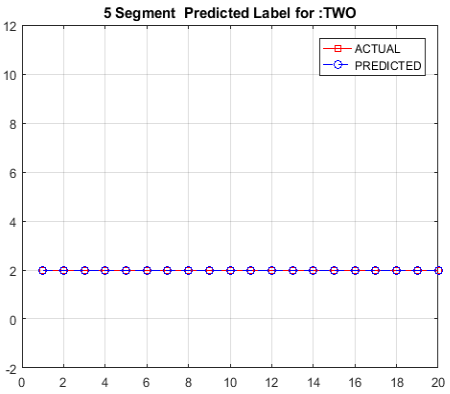
(a)



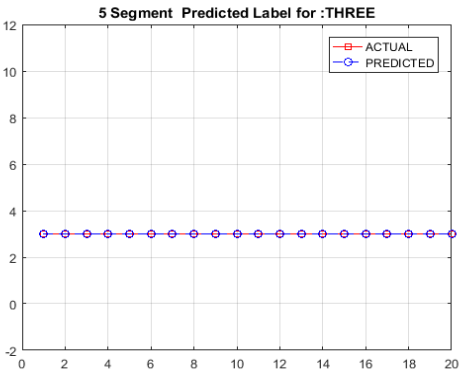
(b)



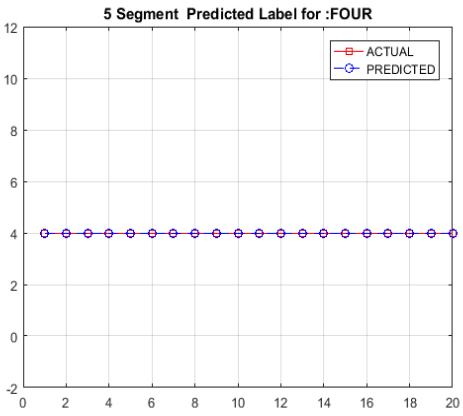
(c)



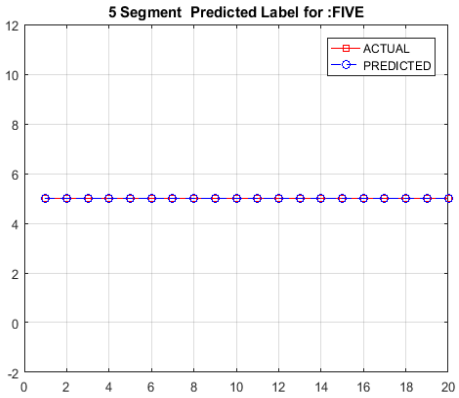
(d)



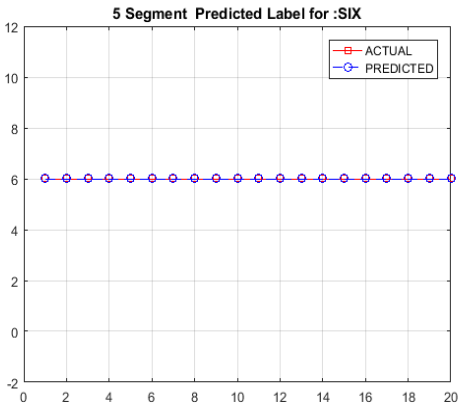
(e)



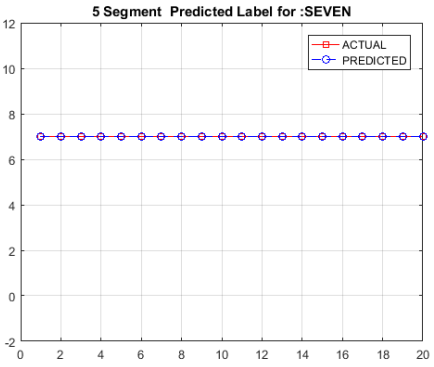
(f)



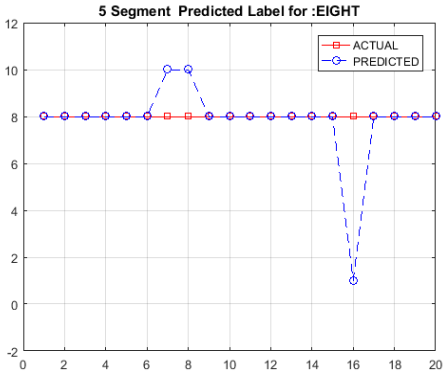
(g)



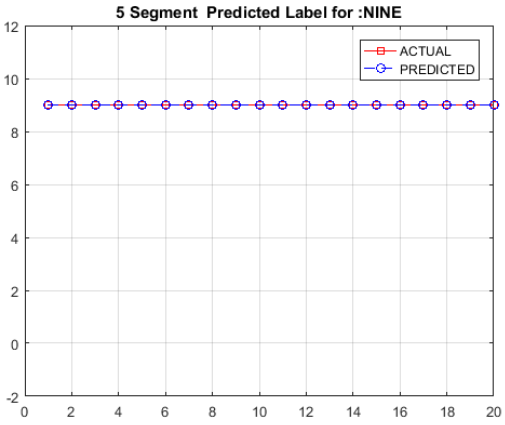
(h)



(i)



(j)



(k)

Figure 5.25 (a)-(k): Relationship between actual and predicted word with 5 segment phoneme separation using KNN

The result shows that five segments of phoneme separation yield accurate results compared to 2, 3 and 4 segments. This gives maximum correlation and occurrence of phonemes.

5.4 Performance of classifier

A confusion matrix is a table 5.14 often used to characterize a classification model's output (or "classifier") on a collection of test data for which the true values are known. Defining the most basic terms now is:

- True positives (TP): These are cases where the outcome was expected yes, but in fact it is no.
- True negatives (TN): These are cases where the outcome was expected no, but in fact it is no.
- False positives (FP): These are cases where the outcome was expected yes, but in fact it is yes.
- False negatives (FN): These are cases where the outcome was expected no, but in fact it is no.

Table 5.14: Confusion Matrix

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Sensitivity = true positive fraction

= 1 – false negative fraction

= TP / (TP + FN)

Specificity = true negative fraction

= 1 – false positive fraction

= TN / (TN + FP)

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Table 5.15 below shows the results of 5 segment phoneme separation using a positive position search algorithm.

Table 5.15: Result of 5 segment phoneme separation

Digit	Correctly Predicted (out of 20)	Not Predicted	Accuracy (% result)
0	19	1	95
1	20	0	100
2	20	0	100
3	20	0	100
4	20	0	100
5	20	0	100
6	20	0	100
7	20	0	100
8	17	3	85
9	20	0	100
10	20	0	100
		Average Accuracy:	98.1818

The above table 5.15 shows that using 5 segment phoneme separation out of 220 samples, 216 samples are correctly predicted.

The following Figure 5.26 represents the performance of k-NN classifier without phoneme separation using confusion matrix

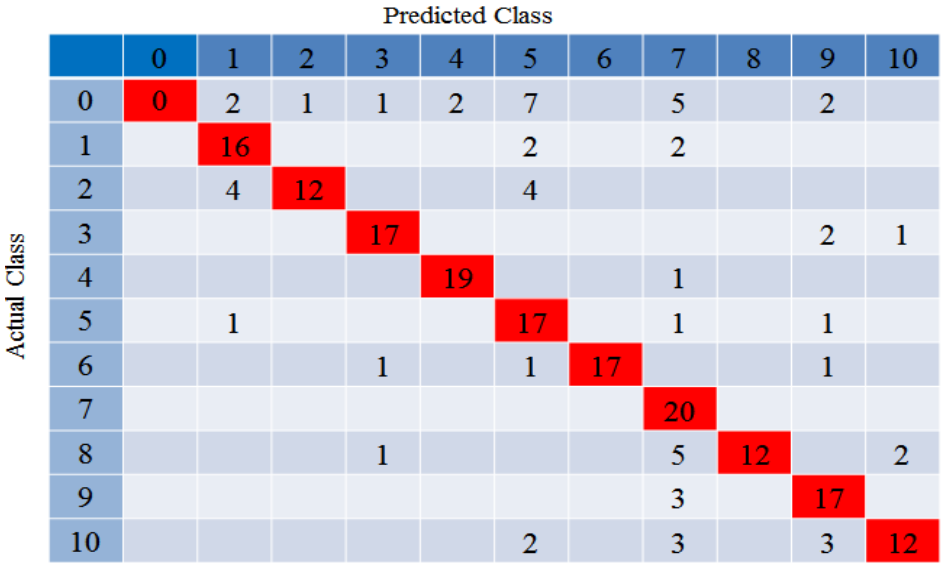


Figure 5.26: Confusion Matrix(Non Phoneme Separation)

Count of correctly predicted words is represented by diagonal elements of confusion matrix. Out of 220 test samples 157 samples are predicted correctly and 63 samples are predicted incorrectly.

The following Figure 5.27 shows the performance of k-NN classifier with phoneme separation using confusion matrix.

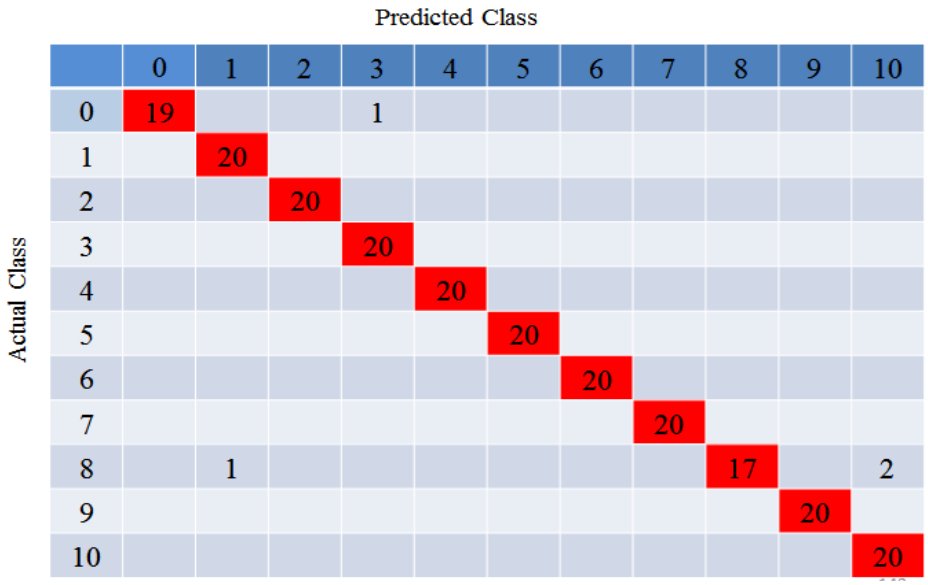
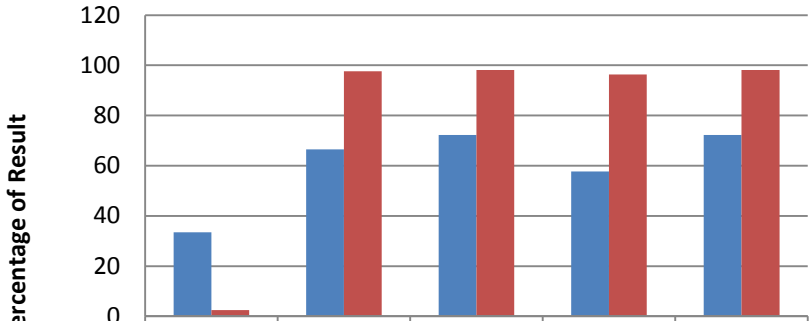


Figure.5.27: Confusion Matrix(Phoneme Separation)

Count of correctly predicted words is represented by diagonal elements of confusion matrix. Out of 220 test samples 216 samples are predicted correctly and 4 samples are predicted incorrectly.

Figure 5.27 offers a comparison of k-NN classifier output with and without the technique of phoneme separation.



	ERR	Accuracy	Sensitivity	Specificity	F-score
Without Phoneme separation	33.52	66.48	72.27	57.64	72.27
With Phoneme separation	2.42	97.58	98.18	96.4	98.18

Figure 5.28: Performance of k-NN classifier

It is observed that with the phoneme separation technique, the device decreases error rate and gives better accuracy sensitivity specificity, and F-score. K-NN classifier together with MFCC as a feature extraction technique and a positive position search algorithm yields very good results for people with articulatory disabilities.

After finding correlation and occurrence of phoneme, the word is formed by concatenation and thus the correction of the word for articulatory handicapped people has been done.

5.5 Summary

The impact of features of speech on system performance was examined in this chapter. It also examined the impact of selecting the right classifier. This study provides insight into the performance of the system with and without the separation of phonemes. It is investigated that the performance of the system is improved with the separation of the phoneme compared to without the separation of the phoneme. In addition, this chapter discussed the simulation results obtained by implementing various algorithms with and without phoneme separation. A comparative analysis with and without a phoneme separation analysis found that 5-segment phoneme separation is an optimal choice for speech recognition and rectification systems for articulatory handicapped people**.**

Chapter 6

CONCLUSION AND FUTURE SCOPE

This chapter presents the conclusions of the undertaken research work which is discussed in the preceding chapters and the further research avenues are listed out.

6.1 Conclusion

There has been a lot of research in the field of speech recognition but still the speech recognition systems till date are not hundred percent accurate. This research work has been attempted to implement the speech recognition and rectification system especially for articulatory handicapped people.

The thorough literature survey about speech recognition for articulatory handicapped people has been carried out. The papers presented by the researchers in 1970, 1972 to 2018. The researchers have work on speech recognition but not on speech rectification for articulatory handicapped people. The proposed system focused on the speech recognition as well as rectification. The standard database for articulatory handicapped people was not available so in this work the database of different people who are suffering from articulation problem recorded using RODE NT1 MIC microphone and NUENDO 4 software. The main challenge in carrying out this analysis is to collect data from people with articulation difficulties. As they suffer from different discrepancies of speech, recording the data is very tedious and time-consuming.

Literature survey reveals a lot of work is being done on this, but there is no perfect system. The quality of the system depends on the various factors such as disorder frequency, environmental conditions, features used and the perfect decision logic circuitry. It is therefore very important to study in depth the various extraction methods, feature selection and correct classification. Also the causes of various speech disorders were studied in order to find a solution for a particular disorder. This study focused on articulation-related problems. The listener does not understand the words of a person with an articulation disorder because the words or sound cannot be properly constructed. The disorder can be caused by physical conditions such as

cleavage of the palate, a disease that causes trouble in making sounds / words, or loss of hearing, or other mouth and cerebral palsy issues. The database related to one of the above issues was considered in this study. Examples of articulation errors include replacing one sound with another (for example, saying ken for ten), or leaving out sounds (for example, tree instead of three) or adding sounds to words ("pinanio" for "piano"). A peculiar change in the letters "s" and "z" is related to the listing problem. The lisping person changes the sounds with "ch" ("seven" sounds like "cheven").

Extraction of features is one of the stages of identification of speech. Work has been done in this research to determine the most efficient method of extraction of features. The literature survey discusses various techniques for extraction of functionality. Among these various techniques, the most commonly used extraction techniques such as LPC, RASTA-PLP and MFCC have been studied in depth. In this research, four techniques were used for the extraction of features: Linear Predictive Coding (LPC), Relative Spectral Perceptual Linear Prediction (RASTA PLP) and Mel Frequency Cepstral Coefficients (MFCC). After studying each of these methods, it was found that they have their own advantages and disadvantages and are all used for various purposes. Research shows that Mel Frequency Cepstral Coefficients have better accuracy and is the most reliable technique than other techniques such as LPC and RASTA PLP for extraction of properties. In this research, the MFCC extraction technique performs well among all techniques since it is perceptually motivated and has non-linear behavioral characteristics. MFCC output also depends on the different parameters such as pre-emphasis filter order, frame width, frame overlap, window type, and number of features choice. The effect of all these parameters was examined in order to determine the correct set of parameters used for the MFCC.

In this research next part was selection of classifier. In order to recognize the correct word, the Classifier plays an important role in this process. There are a lot of choices available to do the same. Classifier selection is mainly based on factors such as complexity, time requirement, memory requirement, flexibility, word error rate, specificity, accuracy and many others. In this analysis, tests were conducted on various classifiers such as Minimum Euclidean distance, SVM, HMM, ANN and k-

NN. The results of the experiment show that the k-NN classifier gives more accuracy compared to other methods. K-NN classifier quality was also tested by varying the k value and various distance measurements. It has been observed that in this study, Fine k-NN performs well with Euclidean distance. A novel approach known as a positive position search algorithm has been implemented for word rectification. Positive position algorithm is doing well for this framework. Using this research's novel approach, speech recognition and correction with phoneme separation technique using MFCC and k-NN classifier achieved 98.81 percent accuracy that is better than word prediction without phoneme separation.

6.2 Future Scope

There remains a massive scope to further enhance the performance of algorithms, developed and implemented in this research. Some of the further work may include:

- The proposed algorithm can be used for continuous speech
- In future the device can be designed which helps to rectify the disorder speech of articulatory handicapped people. So, in real time they can easily communicate with others.

REFERENCES

[1] Corneliu Octavian DUMITRU,Inge GAVAT "A Comparative Study of Feature Extraction Methods Applied toContinuous Speech Recognition in Romanian Language", 48thInternational Symposium ELMAR-2006,Zadar, Croatia, 07-09 June 2006.

[2] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic Recognition of spoken Digits", J.Acoust. Soc. Am., 24(6):637-642, 1952.

[3] H. F. Olson and H. Belar, "Phonetic Typewriter", J.Acoust.Soc.Am., 28(6):1072-1081,1956.

[4] Balaji V, G Sadashiappa, Speech disabilities in adults and the suitable speech recognition software tools-a review,International Conference on Computing and Network Communications (CoCoNet'15), Dec. 16-19, 2015.

[5] Caroline Bowen, "Children's Speech Sound Disorders", Oxford: WileyBlackwell, 2009, ISBN: 978-0-470-72364-7.

[6] Norezmi Jamal, ShahnoorShanta, Farhanahani Mahmud, and MNAH Sha'abani, "Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: A review" AIP Conference Proceedings 1883, 020028 (2017); https://doi.org/10.1063/1.5002046 Published Online: 14 September 2017

[7] S. S. Awad, "The application of digital speech is processing to stuttering therapy," in Instrumentation and Measurement Technology Conference, 1997. IMTC/97. Proceedings. 'Sensing, Processing, Networking", IEEE, 1997, pp. 1361-1367 vol.2.

[8] P. Howell and S. Sackin, "Automatic recognition of repetitions and prolongations in stuttered speech," in Proceedings of the First World Congress on Fluency Disorders, 1995, pp. 372-374.

[9] P. Howell, S. Sackin, and K. Glenn, "Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: I. Psychometric procedures appropriate for selection of training material for lexical dysfluency classifiers," Journal of Speech, Language, and Hearing Research, vol. 40, p. 1073, 1997.

[10] P. Howell, S. Sackin, and K. Glenn, "Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter:

II. ANN recognition of repetitions and prolongations with supplied word segment markers," Journal of Speech, Language, and Hearing Research, vol. 40, p. 1085, 1997.

[11] Y. V. Geetha, K. Pratibha, R. Ashok, and S. K. Ravindra, "Classification of childhood dysfluencies using neural networks," Journal of fluency disorders, vol. 25, pp. 99-117, 2000.

[12] Jun Ren, Mingzhe Liu, " An Automatic Dysarthric Speech Recognition Approach using Deep Neural Networks," International Journal of Advanced Computer Science and Applications, Vol. 8, No. 12, 2017.

[13] Seyed Reza Shahamiri⇑ , SitiSalwahBintiSalim , " Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach," Advanced Engineering Informatics 28 (2014) 102–110

[14] G. Jayaram, K. Abdelhamied, "Experiments in dysarthric speech recognition using artificial neural networks", J. Rehabil. Res. Dev. 32 (1995) 162–169

[15] A. Czyzewski, A. Kaczmarek, and B. Kostek, "Intelligent processing of stuttered speech," Journal of Intelligent Information Systems, vol. 21, pp. 143-171, 2003.

[16] I. Szczurowska, W. Kuniszyk-Jozkowiak, and E. Smolka, "The application of Kohonen and Multilayer Perceptron Networks in the speech nonfluency analysis," Archives of Acoustics, vol. 31, p. 205, 2006.

[17] I. Świetlicka, W. Kuniszyk-Jóźkowiak, and E. Smołka, "Artificial Neural Networks in the Disabled Speech Analysis," in Computer Recognit

[18] Nayak, J., Bhat, P.S., Acharya, R., Aithal, U.V. "Classification and analysis of speech abnormalities." ITBM-RBM 26, 319–327 (2005)

[19] PariaJamshid Lou, Peter Anderson, Mark Johnson, "Disfluency Detection using Auto-Correlational Neural Networks," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4610–4619 Brussels, Belgium, October 31 - November 4, 2018. Association for Computational Linguistics

[20] M. Wiśniewski, W. Kuniszyk-Jóźkowiak, E. Smołka, and W.Suszyński, "Automatic Detection of Prolonged Fricative Phonemes with the

Hidden Markov Models Approach," Journal of Medical Informatics & Technologies, vol. 11/2007, 2007.

[21] E. Nöth, H. Niemann, T. Haderlein, M. Decher, U. Eysholdt, F.Rosanowski, and T. Wittenberg, "Automatic stuttering recognitionusing hidden Markov models," 2000.

[22] Fangxin Chen' and AleksandarKostov, "Optimization of dysarthric speech recognition," Proceedings - 19th International Conference - IEEE/EMBS Oct. 30 - Nov. 2, 1997 Chicago, IL. USA

[23] KT Mengistu, F Rudzicz, " Adapting Acoustic and Lexical Models to Dysarthric Speech," Canadian Conference on Artificial Intelligence, 2011 – Springer

[24] Frank Rudzicz," Using articulatory likelihoods in the recognition of dysarthric speech" Speech Communication 54 (2012) 430–444 ,Elsevier

[25] Santiago-Omar Caballero-Morales, Felipe Trujillo-Romero, "Evolutionary Approach for Integration of Multiple Pronunciation Patterns for Enhancement of Dysarthric Speech Recognition," Expert Systems with Applications an international journal, published by Elsevier2014

[26] Elham S. Salama, Reda A. El-Khoribi, Mahmoud E. Shoman, " Audio-Visual Speech Recognition for People with Speech Disorders," International Journal of Computer Applications (0975 – 8887) Volume 96– No.2, June 2014

[27] Wiśniewski, M., Kuniszyk-Jóźkowiak, W., Smołka, E., &Suszyński, W. (2007)."Automatic Detection of Disorders in a Continuous Speech with the Hidden Markov Models Approach". Computer Recognition Systems 2, 445–453. doi:10.1007/978-3-540-75175-5_56

[28] D. Le and E. M. Provost, "Improving Automatic Recognition of Aphasic Speech with AphasiaBank," Interspeech 2016, pp. 2681-2685, 2016.

[29] T. Lee, Y. Liu, P.-W. Huang, J.-T. Chien, W. K. Lam, Y. T. Yeung, et al., "Automatic speech recognition for acoustical analysis and assessment of cantonese pathological voice and speech," in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, 2016, pp. 6475-6479.

[30] Mark Hasegawa-Johnson, Jonathan Gunderson, Adrienne Perlman, Thomas Huang "HMM-Based And SVM-Based Recognition of The Speech of Talkers

With Spastic Dysarthria"Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on 3:III - III · June 2006

[31] Carlos M. Travieso , Jesus B. Alonso , J.R. Orozco-Arroyave , J.F. Vargas-Bonilla , E. Noth , Antonio G. Ravelo-Garcia , "Detection of different voice diseases based on the nonlinear characterization of speech signals, " Expert Systems With Applications (2017), doi: 10.1016/j.eswa.2017.04.012

[32] Jianglin Wang, Cheolwoo Jo. "Vocal Folds Disorder Detection using Pattern Recognition Methods." 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. doi:10.1109/iembs.2007.4353023

[33] Chen, W., Peng, C., Zhu, X., Wan, B., & Wei, D. "SVM-based Identification of Pathological Voices." 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. doi:10.1109/iembs.2007.4353156

[33] John Labiak, Karen Livescu."Nearest Neighbors with Learned Distances for Phonetic Frame Classification." INTERSPEECH 2011

[34] LadanGolipourandDouglas O'Shaughnessy. "Phoneme Classification and Lattice Rescoring Based on a k-NN Approach." INTERSPEECH 2010

[35] Ooi Chia Ai, M. Hariharan ,SazaliYaacob, Lim Sin Chee "Classification of speech dysfluencies with MFCC and LPCC features." Expert Systems with Applications 39 (2012) 2157–2165

[36] Li-Yu Hu , Min-Wei Huang, Shih-Wen Ke, Chih-Fong Tsai *"*The distance function effect on k-nearest neighbor classification for medical datasets*"*, Hu et al. SpringerPlus (2016) 5:1304 DOI 10.1186/s40064-016-2941-7

[37] T. LakshmiPriya, N.R.Raajan, N.Raju, P.Preethi, S.Mathini,*"*Speech and Non-Speech Identification and Classification using KNN Algorithm*"*,International Conference On Modeling Optimization And Computing,1877-7058 © 2012 Published by Elsevier Ltd.

[38] K U Syaliman , E B Nababan , and O S Sitompul"Improving the accuracy of k-nearest neighbor using local mean based and distance weight"2nd International Conference on Computing and Applied Informatics 2017 Journal of Physics: Conf. Series 978 (2018) 012047 doi :10.1088/1742-6596/978/1/012047.

[39] [40]Gustavo E.A.P.A. Batista, Diego Furtado Silva,2009“How k-nearest neighbor parameters affect it’s performance.” JAIIO - SimposioArgentino de Inteligencia Artificial (ASAI) pp 95-106.

[40] ZuzanaDankoviˇcová, DávidSovák, Peter DrotárandLiberiosVokorokos.“Machine Learning Approach to Dysphonia Detection.” Article from applied sciences, published in 15 October2018

[41] Ratnadeep R. Deshmukh Dr. Babasaheb Ambedkar Marathwada University “Automatic Speech Recognition Techniques: A Review”

[42] Masaki Honda, “Human Speech Production Mechanism”, selected papers, NTT technical review, vol.1, No. 2, May 2003.

[43] Edmund Blair Bolles, “Speech Circuitry”, a blog on origins of speech posted at http://www.babelsdawn.com, Jul 2009.

[44] Hockett, C. F., “The origin of speech” Scientific American, vol. 203, pp. 88-96, 1960.

[45] Goodall, J., The Chimpanzees of Gombe. Patterns of behavior (Cambridge, MA and London: Belknap Press of Harvard University Press, 1986).

[46] Darwin, C., The Expression of the Emotions in Man and Animals (London: Murray, 1872).

[47] Mc Caffrey, Patrick, CMSD 620 Neuroanatomy of speech, Swallowing and language (Neuroscience on the web, California State university, Chico, Feb 2009).

[48] Tool module: The Human Vocal Apparatus, available at http://thebrain.mcgill.ca/flash/capsules/outil_bleu21.html

[49] Fitch, W. T., “The evolution of speech: a comparative review”, Trends in Cognitive Science 4, pp. 258-267, 2000.

[50] O'Saughnessy D., Speech Communication - Human and Machine (Addison-Wesley, 1987).

[51] Breen A., Bowers E., Welsh W., “An Investigation into the Generation of Mouth Shapes for a Talking Head”, Proceedings of ICSLP 96 (4), 1996.

[52] Rossing T., The Science of Sound (Addison-Wesley, 1990).

[53] Rabiner, L. R. and Schafer, R. W., “Introduction to Digital Speech Processing”, NOW, the essence of knowledge, Foundations and Trends in Signal Processing, vol. 1, No. 1-2, 2007.

[54] [Online] available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[55] Bhabad S. S., G. K. Kharate, “Pitch detection in time, frequency and cepstral domain for articulatory handicapped people”, IEEE International Conference in MOOC , Innovation and Technology in Education (MITE), 2013.

[56] Harish Chander Mahendru , “Quick Review of Human Speech Production Mechanism” International Journal of Engineering Research and Development e-ISSN: 2278-067X, p-ISSN: 2278-800X, Volume 9, Issue 10 (January 2014), PP. 48-54

[57] Gonzalez J, Lopez-Moreno I, Franco-Pedroso J, Ramos D, Toledano D, et al. (2010) TATVS-UAM NIST SRE 2010 System Description. 415-8.

[58] Lera G, Pinzolas M., Neighborhood Based Levenberg–Marquardt Algorithm for Neural Network Training, IEEE Trans Neural Netw. 2002;13(5):1200-3. doi: 10.1109/TNN.2002.1031951.

[59] Deividas Eringis ,Gintautas Tamulevičius ,Improving Speech Recognition Rate through Analysis Parameters, Electrical, Control and Communication Engineering doi: 10.2478/ecce-2014-0009

[60] K. Paliwal and K. Wojcicki, “Effect of Analysis Window Duration on Speech Intelligibility,” IEEE Signal Processing Letters, vol. 15, pp. 785–788, 2008.

[61] S. Kim, T. Eriksson, H.-G. Kang, and D. H. Youn, “A pitch synchronous feature extraction method for speaker recognition,” in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. I–405–8.

[62] I. Ding, “Enhancement of speech recognition using a variable-length frame overlapping method,” in Proceedings of International Symposium on Computer, Communication, Control and Automation (3CA), 2010,pp. 375–377.

[63] Q. Zhu and A. Abeer, “On the use of variable frame rate analysis inspeech recognition,” in Proceedings of IEEE International Conferenceon Acoustics, Speech, and Signal Processing. Proceedings (Cat.No.00CH37100), vol. 3, pp. 1783–1786

[64] Z.-H. Tan and B. Lindberg, “Low-Complexity Variable Frame Rate Analysis for Speech Recognition and Voice Activity Detection,” IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 5, pp. 798–807, Oct. 2010.

[65] L. R. Rabiner and R. W. Schafer, “Introduction to Digital Speech Processing,” *Oundations and Trends in Signal Processing*, vol. 1, no.1–2, 2007.

[66] Lei Xie, Zhi-Qiang Liu, “A Comparative Study of Audio Features For Audio to Visual Cobversion in MPEG-4 Compliant Facial Animation,” Proc. of ICMLC, Dalian, 13-16 Aug-2006.

[67] J. L. Ostrander, T. D. Hopmann, and E. J. Delp, “Speech Recognition using LPC Analysis,” *Robot System Division, College of Ingineering,The University of Michigan*, January, 1982.

[68] R Hynek Hermansky and Nelson Morgon(1994,Oct). “RASTA processing of speech” IEEE transaction on speech and audio processing, Vol 2

[69] Davis, S. Mermelstein, P. (1980) *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*. In IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28 No. 4, pp. 357-366

[70] B. Gold, N. Morgan, Speech and Audio Signal Processing, New York, John Wiley and Sons, 2002.

[71] Freeman J. A, Skapura D. M., 2006. Neural Networks Algorithm, Application and Programming Techniques, Pearson Education.

[72] Economou K., Lymberopoulos D., 1999. A New Perspective in Learning Pattern Generation for Teaching Neural Networks, Volume 12, Issue 4-5, 767-775.

[73] Bhushan C. Kamble, 2016. Speech Recognition Using Artificial Neural Network –A Review, Int'l Journal of Computing, Communications & Instrumentation Engg. (IJCCIE) Vol. 3, Issue 1 (2016) ISSN 2349-1469 EISSN 2349-1477

[74] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed,Academic Press, 1990.

[75] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, 2nd ed,Wiley Interscience, 2001.

[76] C. J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 1998, pp. 121-167.

[77] Lera G, Pinzolas M., Neighborhood Based Levenberg–Marquardt Algorithm for Neural Network Training, IEEE Trans Neural Netw. 2002;13(5):1200-3. doi: 10.1109/TNN.2002.1031951.

[78] Yong hong Yan, "Understanding Speech Recognition Using Correlation-Generated Neural Network Targets", IEEE Transactions On Speech And Audio Processing, VOL. 7, NO. 3, MAY 1999, 350-352

[79] Xian Tang, "Hybrid Hidden Markov Model and Artificial Neural Network for Automatic Speech Recognition",2009, 978-0-7695-3614-9

Organizations involved in the treatment and research on speech impairment in children

1. “The Stuttering Foundation” Since 1947 – A non-profit Organization Helping those who Stutter http://www.stutteringhelp.org/
2. Royal College of Speech and Language Therapists (RCSLT) http://www.rcslt.org/
3. The website of American Speech-Language-Hearing Association http://www.asha.org/
4. Dr. Caroline Bowen's www.speech-language-therapy.com
5. Canadian Association of Speech–Language Pathologists and Audiologists (CASLPA).http://www.caslpa.ca

PUBLICATIONS BASED ON THE RESEARCH WORK

1 Bhabad S. S., G. K. Kharate, “An Overview of Technical Progress in Speech Recognition” International Journal of Advanced Research in Computer Science and Software Engineering 3, March - 2013, pp. 488-497.

2 Bhabad S. S., G. K. Kharate, “Pitch detection in time, frequency and cepstral domain for articulatory handicapped people”, IEEE International Conference in MOOC , Innovation and Technology in Education (MITE), 2013.

3 Bhabad S. S., G. K. Kharate, “Speech Recognition using MFCC and various segmentation Parameters for Deaf People”, ICCCC 2017.

4 Bhabad S. S., G. K. Kharate, “Effect of performance parameters of SVM and k-NN on speech recognition for articulatory Handicapped people”, International Journal for Research in Engineering Application & Management (IJREAM) ISSN: 2454-9150 Vol-04, Issue-04, July 2018.

5 Bhabad S. S., G. K. Kharate, “Parameter Estimation Using HMM For Disordered Speech”, 2018 IEEE Workshop on Complexity in Engineering, October 2018, Italia.

PATENT FILED

Sanjivani S. Bhabad, Indian Patent (Filed), 201621039304, “Speech recognition and correction for articulatory handicapped people”.