ANALYSIS OF SPEECH RECOGNITION SYSTEMS FOR MAN MACHINE INTERACTION

THESIS

submitted in fulfilment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

to

J. C. BOSE UNIVERSITY OF SCIENCE & TECHNOLOGY, YMCA

by

SUNANDA MENDIRATTA
YMCAUST/PH 17/2012

Under the Supervision of

Dr NEELAM TURK
Professor, J C Bose UST, YMCA, Faridabad

and

Dr DIPALI BANSAL
Professor, Delhi Skill & Entrepreneurship University

Department of Electronics Engineering
Faculty of Engineering and Technology
J. C. Bose University of Science & Technology, YMCA
Sector-6, Mathura Road, Faridabad, Haryana, India

March 2022

DEDICATION

This research work is dedicated to my family, who has constantly supported me to pursue the research. My husband, Rohit, has always been inspiring me to continue working towards my goal. With his words of encouragement and advice, he provided moral and emotional support, when I thought of giving up. My children, Rishit and Radhika became independent and did the chores on their own, so that I could work upon the targets. With their love and affection, there was no stopping by.

And most of all, I dedicate this thesis to the Almighty God. For it is Him who gave me the strength, knowledge and Wisdom to achieve what I dreamt of.

DECLARATION

I hereby declare that this thesis entitled **ANALYSIS OF SPEECH RECOGNITION SYSTEMS FOR MAN MACHINE INTERACTION** *by* **SUNANDA MENDIRATTA,** being submitted in fulfilment of the requirements for the Degree of Doctor of Philosophy in ELECTRONICS ENGINEERING under the Faculty of Engineering of J. C. Bose University of Science & Technology, YMCA, Faridabad, during the academic year 2020-2021, is a bonafide record of my original work carried out under guidance and supervision of **Dr NEELAM TURK, Professor, Electronics Engineering Department, J.C. Bose University Of Science And Technology, Faridabad** and **Dr DIPALI BANSAL, Professor, GB Pant Okhla Campus 1, Delhi Skill & Entrepreneurship University, Govt. of NCT of Delhi** and has not been presented elsewhere.

I further declare that the thesis does not contain any part of any work which has been submitted for the award of any degree either in this university or in any other university.

Sunanda Mendiratta
YMCAUST/PH 17/2012

CERTIFICATE

This is to certify that this Thesis entitled **ANALYSIS OF SPEECH RECOGNITION SYSTEMS FOR MAN MACHINE INTERACTION** by **SUNANDA MENDIRATTA,** submitted in fulfilment of the requirement for the Degree of Doctor of Philosophy in Electronics Engineering under the Faculty of Engineering of J.C. Bose University of Science & Technology, YMCA, Faridabad, during the academic year 2020-2021, is a bonafide record of work carried out under my guidance and supervision.

I further declare that to the best of my knowledge; the thesis does not contain any part of any work which has been submitted for the award of any degree either in this university or in any other university.

Dr Neelam Turk
(Supervisor)
Professor
Electronics Engineering Department
Faculty of Engineering
J. C. Bose University of Science & Technology, YMCA
Faridabad

Dr Dipali Bansal
(Co-Supervisor)
Professor
GB Pant Okhla Campus 1
Delhi Skill & Entrepreneurship University
Govt. of NCT of Delhi

Dated:

ACKNOWLEDGEMENT

I wish to record my sincere gratitude to my Supervisors **Dr Neelam Turk** and **Dr Dipali Bansal** for allowing me to work in this area. I thank them for their constant support, motivation, and recommendation. It would never be possible for me to take this thesis to this level without their innovative ideas and their relentless support and encouragement. I am indebted to them more than they know.

I would like to express my gratitude to **Prof. Sushil Kumar Tomar**, Vice-Chancellor, J.C. Bose University of Science and Technology, YMCA, Faridabad, for his motivation and support.

My sincere thanks to **Dr. Munish Vashishtha,** Professor, Department of Electronics Engineering, J.C. Bose University of Science and Technology, YMCA, Faridabad for his thoughtful suggestions and recommendations that helped me widen my research from various perspectives.

I also take the opportunity to thank **my peer scholars, staff and laboratory staff** of the Department of Electronics Engineering at J. C. Bose University of Science and Technology, YMCA for their support, cooperation and encouragement.

Words fail me to express my appreciation to my husband, **Mr Rohit Mendiratta** whose persistent confidence in me has always prompted me to pursue academic activities deeply. I also express my deep sense of gratitude to my in-laws and my parents for the cooperation, encouragement, support and affection they bestowed upon me. I am also thankful to my lovely children, **Rishit** and **Radhika**. Their natural smiles extended to me, relief all through this tiresome endeavour.

Sunanda Mendiratta

Registration No. YMCAUST/PH 17/2012

ABSTRACT

Speech is a basic way of communication between humans but, over the past few decades, it has been the best way to communicate between machines. It is faster and more convenient than the conventional methods of feeding the machines with devices like keyboards, mouses, light pens and touch screens. Nowadays, speech is recognized by a laptop or a mobile and is converted into text using Automatic speech recognition (ASR) systems. Automatic speech recognition (ASR) systems consist of Speech Processing and Speech Recognition. Automatic Speech Recognition is the ability of a machine to recognize speech and convert it to text.

Speech Processing is the study of speech signals and the processing methods of these signals which includes acquisition, filtering, noise removal, analogue to digital conversion, storage etc. The signals are usually processed in a digital representation. Speech processing is, therefore, a special case of digital signal processing applied to the speech signal. The main applications of speech processing are the recognition, synthesis and compression of human speech. Speech Recognition is also called voice recognition and deals with the analysis of the linguistic content of a speech signal and its conversion into a computer-readable format. It is also called speech to text conversion or simply STT.

Automatic speech recognizers can be used to facilitate communication between humans and machines. Speech-based, man-machine interaction is demonstrated in several applications. The applications include voice-mail systems in telephony, hands-free machine operations, communication interfaces for people with special abilities, dictation systems, and translation devices. More recent applications include the voice assistants like Siri iOS (Apple), google voice assistant and Alexa from Amazon. These voice-based applications can perform internet searches and answer your queries. However, there are a certain number of limitations such as language and vocabulary, speaker dependency, noise-free environments and low talking rates. But still, researchers are working in this field to get better results.

The basic idea behind Automatic Speech Recognition (ASR) in the context of Isolated Word Recognition (IWR) has been explored. The goal of ASR is to covert a speech signal into its equivalent text message independent of the device, speaker or environment. It is a problem of pattern recognition in which features are

extracted and a model is used for training and testing. Various Machine Learning techniques are also proposed to produce a robust speech recognition system.

Having a machine to understand fluently spoken speech has driven speech research for more than eight decades. Although ASR technology is not yet at the point where machines understand all speech, in any acoustic environment, or by any person. It is still used on a day-to-day basis in several applications and services. A good speech recognition algorithm must produce an accurate recognition of speech. Various researchers have developed speech recognition systems by using different feature extraction approaches. In the existing work, a speech pre-enhancement method for efficient speech recognition is developed based on matching the recognized text to the original message. The noisy features were extracted by the approximation of some selected criteria. The speech signal was well recognized in the presence of environmental noise and with a low signal-to-noise ratio. However, the accuracy of the recognition is not up to the mark. The words recognized by the existing method are greatly varied through accents, dialects and mannerisms. Those accents, dialects and mannerisms are some of the most important factors in properly recognising a word as well as identifying the difference between other words. An accent means the way of pronunciation of a particular speaker and dialects refer to a variety of languages that is a characteristic of a particular group of the language's speakers. Hence to make a perfect ASR system the accents, dialects and mannerisms must be recognized well. In this work, the aim has been to develop a better automated speech recognition system by overcoming the drawbacks of the existing systems.

The main components of the ASR based on a statistical approach are feature extraction, acoustic models (HMMs), language models and hypothesis search units. The acoustic model typically consists of two parts. The first is to describe how a word sequence can be represented by sub-word units and the second is the mapping from each sub-word unit to acoustic observations. In the language-model rules are introduced to follow the linguistic restrictions present in the language and to allow redemption of possible invalid phoneme sequences. The acoustic and language models resulting from the training procedure are used as knowledge sources during decoding.

The purpose of the present study is to contribute to an understanding of the present-day speech recognition systems by studying the various algorithms being used for

feature extraction and classification of speech/voice input. The main objectives of the research were as follows. After studying the literature, standard stored databases and real-time data were studied and acoustic models have been created. The three main processes of preprocessing, feature extraction and classification were performed on the acquired database. To remove noise, techniques like wiener filtering, spectral subtraction and windowing were used. The analogue voice signals were converted to digital voice signals for further processes like sampling, windowing, and framing. The common features and the statistical features of the voice signal were extracted. And Pattern recognition was performed. Then Artificial Neural Network with Backpropagation was used as a classifier. An attempt was made to recognize human emotional states of happy, sad and neutral during the speech.

In this research, an ASR system is designed for man-machine interaction. For this five approaches have been proposed and they are: i) Isolated word recognition system for speech to text conversion using ANN, ii) Automatic speech recognition using an optimal selection of features based on hybrid ABC-PSO, iii) Fuzzy Based Selection of DWT Features for Automatic speech recognition System for Man-machine Interaction with CS-ANN Classifier, iv) automatic speech recognition by cuckoo search optimization-based artificial neural network classifier, and v) automatic speech recognition system model for news transcripts The proposed system has been implemented in the working platform of MATLAB.

In this work, a robust speech recognition system using novel innovative feature extraction and classification technique has been developed. The following three phases depict the flow of work.

1) Preprocessing,

2) Feature Extraction, and

3) Classification.

In the first phase, the noise level of the recorded audio signal is measured and fragmented using the Discrete Fourier Transform (DFT). In the second phase, the most required features of the audio signal are extracted for better classification. In the proposed method some of the important features like sampling point, pitch measure, and word length are extracted, along with these features an innovative phoneme intelligibility feature is also extracted for the ASR purpose. These are the general features acquired for speech recognition. However, to reduce the

accent, dialects and mannerism level of the recognized signal some more features are required. These variations in a speech signal occur due to the improper prediction of the emotional speech signal. Hence in the proposed system, emotional feature extraction is also considered. The phoneme feature is one of the well suitable features for predicting the emotional speech signal. In the proposed method the optimum level of features is selected from the extracted four features like sampling point, pitch, word length and phenome feature for the training purpose in classification. The common features like word length, sampling point, pitch; and statistical features like mean, standard deviation, skewness, kurtosis and entropy of the voice signal are extracted and Pattern recognition is performed. Another important feature extraction technique involves the MFCC and LPC features. Short term analysis is used to calculate the MFCC of the speech signal. In the last phase, the audio signal is classified based on the extracted features. For classification, one of the best classifier algorithms is used. The extracted Mel features are given to CSO based ANN. The proposed system is implemented in the working platform of MATLAB 2013a. Then the performance is analyzed by comparing the proposed classifier with the other classifier systems.

TABLE OF CONTENTS

LIST OF TABLES

TABLE **Page No.**

LIST OF FIGURES

FIGURE **Page No.**

LIST OF ABBREVIATIONS

Abbreviation	Definition
ABC	Artificial Bee Colony
ACO	Ant Colony Algorithm
AFE	Advanced Front End
AI	Artificial Intelligence
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AR	Augmented Reality
ASL	Automatic Sign Language
ASR	Automatic Speech recognition
BP	Back Propagation
BPF	Band Pass Filter
BPNN	Back Propagation Neural Network
CNN	Convolutional Neural Network
CSL	Computer Speech and Language
CSO	Cuckoo Search Optimization Algorithm
CSR	Command Success Rate
DARPA	Defence Advanced Research Projects Agency
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
DTW	Dynamic Time Warping
DWT	Discrete Wavelet Transform
EMG	Electromyography
ETSI	European Telecommunications Standard Institute
FB	Filter Bank
FCBF	Fast Correlation Based Filter
FE	Feature Extraction
FISTA	Fast Iterative Shrinkage Thresholding Algorithm
FN	False Negative

Abbreviation	Definition
FNR	False Negative Rate
FP	False Positive
FPGA	Field Programmable Gate Array
FPR	False Positive Rate
FSM	Finite State Machine
GMHMM	Gaussian Mixture Hidden Markov Model
GPS	Ground Positioning System
HCI	Human-Computer Interaction
HMM	Hidden Markov Models
HTK	Hidden Markov Model Tool Kit
HTML	HyperText Markup Language
IWR	Isolated Word Recognition
LASSO	Least Absolute Shrinkage and Selection Operator
LBG	Linde-Buzo-Grey algorithm
LDCIL	Linguistic Data Consortium for Indian Languages
LM	Language Model
LME	Large Margin Estimation
LPC	Linear Predictive Coding
LPCC	Linear Prediction Cepstral Coefficient
LVCSR	Large Vocabulary Continuous Speech Recognition
MATLAB	Matrix Laboratory
MCNR	Multi-Channel Noise Reduction
MEDC	Mel-energy Spectrum Dynamic Coefficients
MFCC	Mel Frequency Cepstral Coefficients
MLP	Multi-Layer Perceptron
MMI	Man-Machine Interaction
MMSE	Minimum Mean Square Error
MRE	Minimum Rank Error

Abbreviation	Definition
MSE	Mean Squared Error
MVDR	Minimum Variance Distortionless Response
MVN	Mean and Variance Normalization technique
NN	Neural Network
NNLM	Neural Network Language Model
ORF	Oral Reading Frequency
PAC	Phase Auto Correlation
PLDA	Probabilistic Linear Discriminant Analysis
PLP	Perceptual Linear Prediction
PNCC	Power Normalized Cepstral Coefficients
PSO	Particle Swarm Optimization
RAPT	Robust Algorithm Pitch Tracking
RATS	Robust Automatic Transcription of Speech
RBF	Radial Basis Function
RBM	Restricted Boltzmann Machines
RNN	Recurrent Neural Network
ROC	Region of Convergence
SANN	Self-Adjustable Neural Network
SESC	Simulated Emotion Speech Corpus
SID	Speaker Identification
SNR	Signal to Noise Ratio
SOCP	Second-Order Cone Programming
SR	Speech Recognition
SRE	Speech Recognition Engines
SSR	Silent Speech Recognition
SSVM	Structured Support Vector Machines
STT	Speech To Text
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TTS	Text To Speech

Abbreviation	Definition
VOCODER	Voice Coder that analyses and synthesizes the human voice signal
VODER	Voice Operating Demonstrator
VSSNLMS	Variable Step Size Normalized Mean Square
VTS	Vector Taylor Series
WFST	Weighted Finite-State Transducers
WRR	Word Recognition Rate
ZCPA	Zero-Crossing Peak Amplitude

CHAPTER 1.

INTRODUCTION

1.1 OVERVIEW

The fundamental mode of communication among humans is speech and in the case of machine-human interface; verbal language has been believed as the natural method. When communication with machines is carried out, it is very difficult and slow-moving in magnitude when realized via keyboards, mouses and other devices. Thus, speech feed-in is an important constituent to making this communication easily accessible. Also, humans see speech as a great source of information. Therefore, persons who are not literate or have vague about computers can easily access computers by employing speech instructions[1]. Even people with some physical disability who are not able to type or click with their hands can use their speech to operate the computer. Even people who are proficient in operating computers can speed up data entry, sending emails and other documents using the speech input methods. Furthermore, this mode of operation possesses many advantages. For example, while driving, the hands of the driver are busy steering the driving wheel and he cannot type on his mobile. In such a case speech is a good input option. GPS (Global Positioning System) is an example of a speech-based system being used. Another example is speech-enabled dialling, where the user can just ask the device to call a particular person, without dialling his number.

The common and efficient means of communication among humans is through 'speech'. To process speech means to extract useful information from it, processing includes the implementation of electric signals on the acoustic pressure waves collected from human vocalization and applying mathematical analysis to it. The field of processing speech involves the natural operation of analysing speech, coding, augmentation, synthesis, and recognition. Analysis of speech is the study of its creation mechanism to make a mathematical model of physical phenomena. Speech coding aims to keep information about specific speech parameters for later retrieval. The method of refining precision and quality of speech which is noisy utilises various algorithms is recognized as speech enhancement [2]. Producing artificial human speech using coded information is known as the synthesis of speech. The method of inverse synthesis is the capability of a program or machine to classify the linguistic contents mixed up in the speech signal.

Speech production is a very complex procedure. The normal breathing mechanism allows the air to enter the lungs. The tensed vocal chords inside the larynx start vibrating as air from the lungs pass through it. The airflow is sliced into quasi-periodic pulsations which are then tuned in frequency while travelling through the pharynx, the mouth hole, and feasibly nasal hole.

SR study has a past of 80 years. The past 60 years of study in Automatic Speech Recognition and synthesis of speech via machine has gained considerable attention for purposes like technical interest about the techniques for machine-driven apprehension of human speech potential, urge to mechanise functions demanding interface of man-machine. The very first ASR system based on acoustics phonetics was constructed in 1950. These systems involved spectral assessments, using spectrum scrutiny and pattern matching to perform recognition choices on tasks for instance vowel recognition. Some systems employed filter bank analysis to extract spectral details of speech signals [3]. New techniques started emerging in the 1960s like speech segmentation, Zero-Crossing Analysis (ZCA) and dynamic time arrangement and tracing concept. A major revolution was brought in the 1970s in the field of ASR. The Dynamic Time Warping (DTW) technique was developed for isolated word recognition. Earlier Linear Predictive Coding (LPC) was involved in speech coding but later utilized for SR systems formed on LPC spectral parameters. Later in the 1970s IBM developed a system for a large vocabulary SR system, which ended up with significant influence on the ASR arena. AT & T Bell Labs also initiated to create speaker liberated SR schemes by employing clustering algorithms for generating speaker liberated patterns. The connected word recognition system was developed on an idea relying on algorithms that integrated isolated words for recognition. After the mid-1980s researchers started extensively using Hidden Markov Models (HMM) for ASR. Then the concept of Neural Networks was presented in the late 1980s for classifying the signals.

The research is still on in this field making it an attractive or hot topic for researchers. Usually, an SR system attempts to recognise the elementary unit in several linguistic, words or phonemes which can be gathered into a transcript. The possible implementation of the ASR system is in automatic call routing, computer language to the transcript, and translation of machine language. Automatic speech recognition is an interdisciplinary field that extracts theoretical information from physics, mathematics, and engineering. Machine

learning, random processes, information theory, signal processing, pattern recognition, linguistics and psychoacoustics are the essential processes included in ASR systems.

1.2 MOTIVATION

These days, SR is widely used because of its multi-purpose and extensive variety of applications. SR is used in nearly so many areas of life for example in the weather forecast, military, mobile applications, agriculture, datum inquiry, healthcare, computerized speech transformation, command and control, phone directory help, robotics, transcription, office dictation devices, computer games, etc. Furthermore, conveying data to the PC via the keyboard is a lengthy process on the other hand data get transferred very quickly through spoken language.

At present, nearly all technologies associated with man-machine communication are restricted to those who can read, write and have the least understanding of knowledge, mainly English. Therefore, spoken language interactions with computers topic have attracted and captivated many speech scientists and engineers. On the other hand, this also prerequisites a certain understanding of the language. If this human-machine interaction becomes possible in native languages, even people who are not literate can benefit from IT offered advantages. Today, folks have understood the requirement for an effective humanoid-machine interface system based on Indian languages [4]. This motivated researchers to enhance the correctness of the existing speech handling methods in these languages to enable convenient communication schemes. With the development of an efficient speech recognizer, a normal humanoid-machine interaction can be realized. Any person can interact with a computer, who can speak, without requiring devices or specific tools. Speech recognition is becoming a hot topic to explore for researchers as this field holds so much potential, even though it is predicted that voice identification might progress as the prime user interaction in the forthcoming years.

1.3 PROBLEM STATEMENT

The key goal of Automatic Speech Recognition is to come up with an accurate and highly efficient technique to identify human speech through diverse algorithms performed on a PC. Automatic Speech Recognition is a serial pattern recognition task that contains multiple classes. The significant features are drawn from the speech signals which set up the feature vector space and these are categorised as the equivalent speech samples. Speech

recognition (SR) becomes a difficult task as it includes several challenges in forming ASR. SR is a hard task as several issues affect the execution of an SR system. The major limits are the style of speaking, unpredictability in the channel, speaker, and sex of the speaker, spoken language, speech speed, regional and social dialects, and uncertainty in speech, pitch, phonetic identity, microphone, and the existence of background noise.

1.4 OBJECTIVES OF RESEARCH

The goal of ASR is to convert a speech signal into its equivalent text message independent of the device, speaker, or environment. After studying the extensive literature available on the proposed topic, standard stored speech databases and real-time data are studied, and acoustic models are created. The research findings of the researchers in the field of speech recognition are studied.

The objectives of the research are:

•To analyze standard database and real-time data for speech recognition and convert the analog voice signals to digital voice signals for further processes like sampling, windowing and framing.

•To perform preprocessing of the acquired speech signals, and remove noise using techniques like wiener filtering, spectral subtraction, and windowing; and perform word detection and background noise removal.

•To extract the common features and the statistical features of the voice signal and perform classification of the test signal.

•To use Artificial Neural Network with Backpropagation as a classifier and attempt to increase the efficiency of Automatic Speech Recognition Systems.

1.5 SPEECH SIGNAL

Speech comprises a sound series and the exchange of sounds among humans function as information. Speech signal travels as a longitudinal wave in a medium like water or air. The speed of speech signals varies depending upon the medium density [5]. Generally, the change in the magnitude of air pressure corresponding to a speech signal is represented in a graph as a function of time. This type of graph is known as a speech waveform or speech pressure waveform.

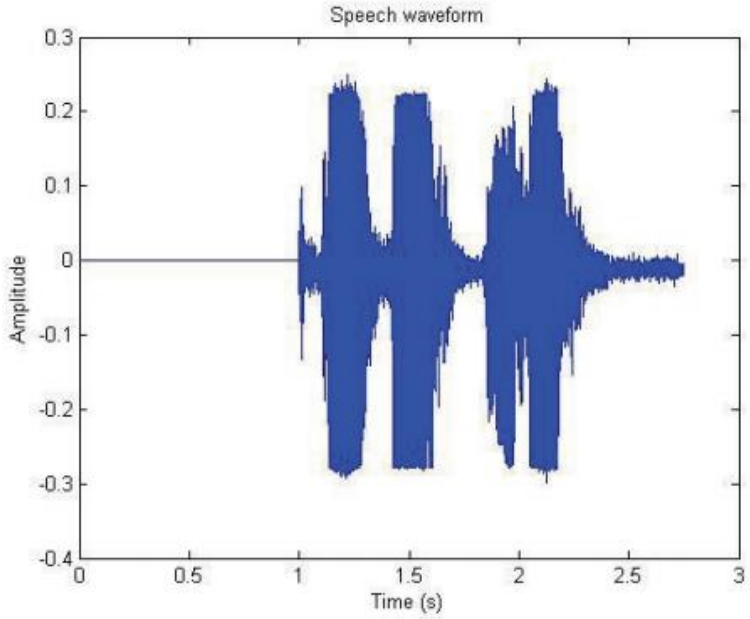


Figure 1.1: Speech waveform

The sinusoid is the basic type of sound that can be defined when defining the characteristics of sound waves. As it is a basic and simplest type of sound wave it is very convenient to start with. So, the properties of the sinusoids are described before discussing the properties of the speech signal.

1.5.1 Properties of Sinusoids

A simple smooth up and down trailing of a time waveform forms a sinusoid. 3 types of evaluation can be drawn from a sinusoid that can describe the outline of a waveform when they are considered together. These evaluations are listed below:

(i) Amplitude: Extent of the waveform movement above and below the midline of a sinusoid. This extent tells about the energy contained in the sound wave and it corresponds to the loudness of the sound wave. Amplitude can be measured in many ways; like in terms of units of pressure as it shows pressure variation in the air.

More often amplitude is measured on a logarithmic scale called decibel (dB) which is comparative to the sound. The decibel scale graph is the same in the way that an individual perceives high volume and hence be found to be a very useful form of measure.

(ii) Frequency: It can be defined as the number of cycles the sinusoid makes per second. A complete cycle consists of oscillation starting from zero or from the midline to the

maximum value, then crossing the midline down to the minimum value and back to the zero or midline. S.I unit to measure frequency is Hertz (Hz) or cycles/second.

(iii) Phase: The position of the starting point of the sinusoid is termed a Phase. The phase of those starting at the maximum is said to be zero while a phase of π radians is for those beginning at the least value. Perceiving the phase of a sinusoid is not possible but the relative changes in the phase between two signals can be detected, as the mind determines the origin of a sound based on different phases heard in both ears forms the basis of binaural hearing.

1.5.2 Properties of the Speech Signal

Significant energy is present in speech ranging from zero frequency up to around 5 kHz. The speech signal properties change remarkably as a function of time. The idea of time-varying Fourier depiction is utilized to study the spectral characteristics of the speech signal. Though, the temporal properties of speech signals like correlation, energy, ZCA, etc. are supposed to remain invariable for a small duration of time, which means their properties remain consistent or fixed for the small-time duration [6]. So, the signal is partitioned into several short-duration blocks using the hamming window, to take the Fourier transform of the speech signal. Some of the important properties of speech signals are summarized as follows.

(i) Bandwidth: The bandwidth of the speech signal is much higher than 4 kHz. Still there exist a substantial portion in the spectrum of energy for high-value frequencies and even very high frequencies for the fricatives. Though, it is known that from using the (analogue) phone, it appears that the speech signal encloses the whole information required to recognize a human voice within a bandwidth of 4 kHz.

(ii) Fundamental Frequency: The usage of vocalized stimulation for the speech signal results in a pulse wave, which is termed the elemental frequency. The signal has an elemental frequency ranging from 80 Hz to 350 Hz and is periodic. While voiced excitation is normally used in the case of articulating vowels and some of the consonants, unvoiced excitation (noise) is used for spirant (e.g., /s/ as in mess, or /f/ as in fish). Generally, no elemental frequency can be identified in such instances. Conversely, the signal has a very high zero-crossing rate. Plosives (like /p/ as input), that show fleeting stimulation can be

primarily detected in the speech signal by inspecting for the small quietness necessary to build up the airborne pressure afore the plosives are busted out.

(iii) Top notches in the Spectrum: The vocal tract offers a characteristic spectral figure to the speech signal after passing through the glottis. If anyone untangles the vocal tract into a straight tube, it is known that the tube shows a resonance at some frequencies. Those frequencies are termed formant frequencies. The frequency of the formants (especially of the 1st and 2nd formant) changes conditional on the figure of the vocal tube (tube diameter changes laterally the tube), and therefore the vowel being articulated is characterized.

(iv) The Power Spectrum Envelope Decreases with Increasing Frequency: The power spectrum of a pulse sequence from the glottis decreases towards higher frequencies at a rate of -12dB per octave. A high pass feature with +6dB per octave is shown by the emission characteristics of the lips show. Thus, an overall decrease of -6dB per octave is the result. Thus, with increasing frequency, there occurs a decrease in the envelope of the power spectrum of the signal.

1.5.3 Human Auditory System

It is necessary to know the auditory system of humans to layout the speech-based interaction system. In any verbal conversation, firstly, the thought to speak comes into the mind of the person then the person speaks which is called speech [7]. Thought in the mind is converted into words, sentences and phrases following the language rules. If seen from the physiological communication point of view, electric signals are generated in the brain that travels through motor nerves. As a result, the vocal tube and vocal cords muscles get activated as soon as they receive the electric signals. Consequently, this activation creates pressure variation in the vocal tube and correspondingly at the lips, which leads to the production of sound wave that travels into space. Lastly, the listener in the verbal conversation in his brain recognizes the speech and understands it.

1.5.4 Speech Production

Human speech is produced by complicated interfaces amongst the diaphragm, lungs, mouth, throat, and nasal passages. Phonation, resonation, and pronunciation are the procedures that handle or govern the production of speech. When air pressure is converted into sound by using vocal cords is known as Phonation. When a few frequencies are highlighted in the vocal tube by resonances is known as Resonation.

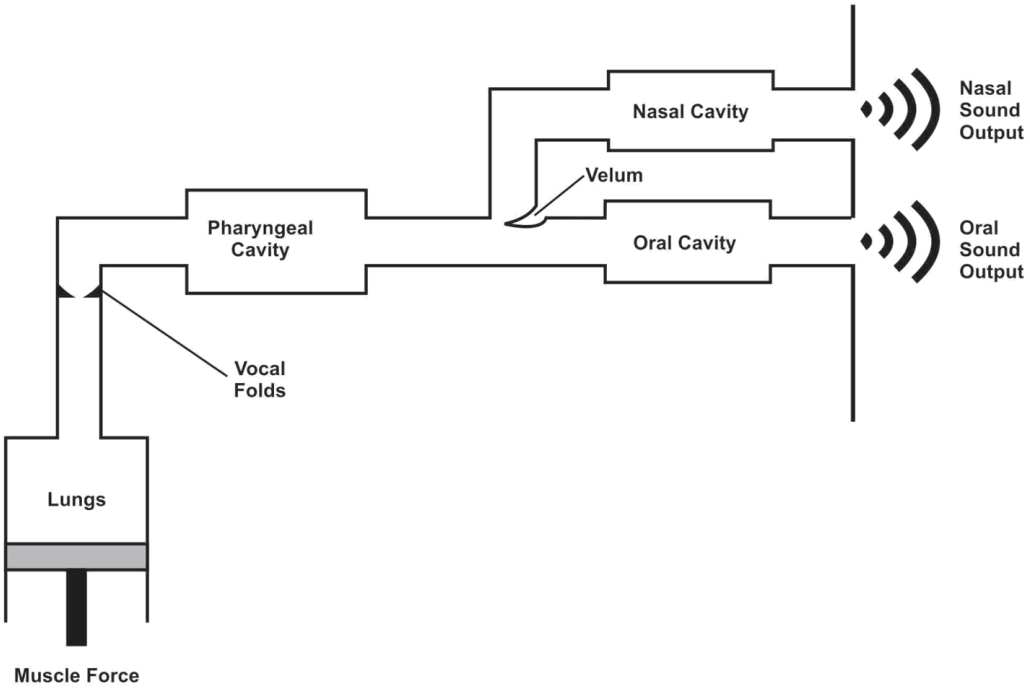


Figure 1.2: Representation of speech production Mechanism[8]

When these vocal tube resonances lead to the production of different sounds is known as pronunciation /Articulation.

Speech production takes place through the biological process of a soundbox or larynx, which exist in the gullet of humans. This is illustrated in figure 1.2. The fibres present inside the larynx, or the soundbox starts vibrating whenever the air passes through them. The larynx is attached to the lung via the windpipe, and the lungs are the main organs that push air down the larynx. The vocal cords or fibres present inside the larynx are proficient in vibrating at every frequency. In some cases, the audible range varies from 20 Hz up to about 11000 Hz. The high-value frequencies are commonly seen in kids' and females' voices whereas the low-value frequencies are seen in males.

Regarding Signal Processing, the larynx and the vocal tract can be treated as the source and the filter respectively. The vocal tract comprises the vocal tube, the mouth hole & the nasal passage, which filters the sound thus inducing to echo of certain frequencies and anti-resonate other than these frequencies to originate a distinct sound from the lips, which is known as speech.

The length and shape of the vocal tract don’t change continuously but remain constant over a scale of a millisecond (ms). Therefore, the properties of the filter are constant over the

scale of milliseconds. Hence on a small scale, a person's voice has a fixed character and on a large scale has a versatile character.

For sound production, the air firstly passes through the lungs, then goes to the glottis, from there reaches the throat and finally to the mouth [9]. Based on speech sound articulated by the speaker the speech signal can be stimulated in three ways:

(i) Stimulated by voice: The pressure is exerted by air onto the glottis, which is initially in the closed condition, which systematically opens and closes it thereby creating a cyclic train of pulses (triangle-shaped). The range of this elemental frequency is 80Hz to 350Hz.

(ii) Stimulated by no voice: when the glottis is not in the closed condition, the air directly travels through a narrow opening in the mouth or throat. Consequently, leads to turmoil which causes noise generation. The spectral shape of the noise signal depends on the narrowness of the opening.

(iii) Transient excitation: The air pressure is created because of a close in the mouth or throat. The pressure of air falls instantly through the unexpected opening of the closed mouth.

In several cases, the speech signal gets stimulated in a combination of these three excitations. The spectral properties of the speech signal depend on the vocal tract shape. The spectral figure of the speech signal changes as the shape of the vocal tract changes due to which different sounds are pronounced.

The next paragraphs will explain the speech production model and speech perception.

1.5.5 Speech Production Model

This model functions as mentioned following. A pulse generator models voiced excitation by which it causes a train of triangle shape pulses whose spectrum is defined by P (f). A white noise generator models the unvoiced excitation with a spectrum, N (f). The amplitude of the signal generated by impulse generator (v) and the noise generator (u) can be modified by mixing voiced and unvoiced excitation. The outcome from each generator is then added and becomes the input of the box which models the vocal tract and gives the spectral shape to the signal with the transmission function H (f). R (f) models the features of the emitted signal from the lips.

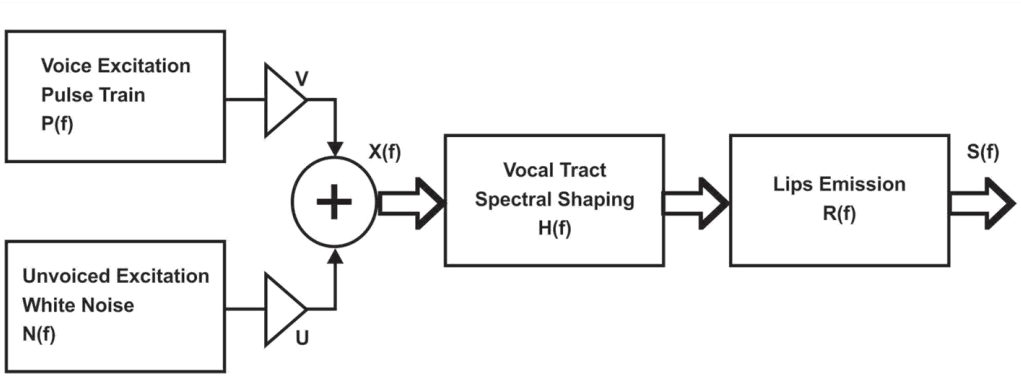


Figure 1.3: A model demonstrating the production of Speech

Therefore, the speech signal spectrum S (f) is specified as:

$$S(f) = (v.P(f) + u.N(f)).H(f).R(f) = X(f).H(f).R(f) \quad (1.1)$$

This speech production model exhibits the below-mentioned features to affect the speech sound:

- P(f) gives the fundamental frequency
- H(f) defines the signal's spectral shape
- v and u decide the amplitude of the signal
- v and u determine the excitation of the signal based on the mixing of voiced and unvoiced excitation

The speech signal is defined by the above-mentioned technical terms. The features mentioned above must be calculated from the time signal to carry out the recognition of speech and then must be despatched to the speech recognizer [10]. The variation in the spectral shape of the signal in the time domain gives important information to the speech recognizer.

1.5.6 Speech Perception

A symbolic diagram of the human ear displaying the three different sound processing units comprises: the external portion of the ear includes the Pinna, which collects the sound and transmits it to the central portion of the ear through the external canal; the central part of ear begins at the tympanic membrane (eardrum) and together with three very small bones namely, the Malleus/hammer, the Incus/the anvil and the stapes/stirrup, which converts audio waves to mechanical pressure waves; and ultimately, the innermost part of the ear,

comprising the cochlea and the set of neural links to the audio nerve, which passes the neural messages to the brain.

1.6 SPEECH RECOGNITION

Speech Recognition (SR) is a special case of pattern recognition. In the SR process, the speech signal is transformed into a word sequence with the help of an algorithm realized as a program. SR aims to give a device the ability to "hear," get it," and "act upon" vocalized data. Two main processes of speech recognition include the training phase and the testing phase [11].

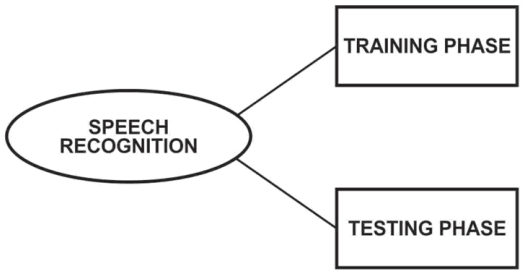


Figure 1.4: Speech Recognition phases

In both phases, the classification of the signal is based on extracted features of the speech signal. The training phase determines the features of the classification model based on training data or class examples. For the recognition phase, the testing data features are compared with the feature set extracted in the training phase. The tested data is assigned to the model which matches exactly or maximum with it

1.6.1 The Concepts

An observation is a scheme that ascribes labels to the ongoing events in the surroundings. If the assigned labels fit sets without a metric distance it is believed classification is the consequence of the observation and the labels match one of several sets. If conversely, the sets are linked through a metric, it implies the outcome is an estimation and the labels belong to a metric space. Rendering to these descriptions, this study aims to develop an observer that defines air pressure waves using the labels enclosed by some inscribed language. This procedure is called classification, as these labels are not related to a metric.

Why has speech recognition become so hot a topic for researchers and attracted funding? It is possible to create a natural man-machine interaction by producing an effectual speech

recognizer. Here ‘natural’ signifies something instinctive and simple to use for a person, or it can be said, a procedure that only requires the natural capabilities of the user rather than using some special tools. This type of system can be understood by any human who can speak and enables the extensive use of machines like computers [12]. This likely assures enormous economic advantages to those who acquire expertise in the techniques required to resolve the problem and describes the attraction towards this arena from the past decade.

By creating an improved machine for speech recognition through better speech-generating methods and natural language systems, it may become possible to develop computational applications that do not require a screen or a keyboard. This may also let unbelievable shrinking/miniaturization of existing systems to enable the construction of small-sized smart devices that can communicate with humans simply by using speech. Such kind of device/machine exists at Carnegie Mellon University, namely, the JANUS system that performs real-time speech recognition and translation of language between Japanese, English and German. An impeccable model of this scheme could be commercially installed to let upcoming clients of separate countries communicate without stressing over their lingual differences. Financially such type of device would cost very high. Phonemes and inscribed words obey cultural protocols. The speech recognizer cannot carry out classifications on its own. It must obey the cultural protocols that describe the selected language. Thus, it states that a speech recognizer must work according to cultural rules. It is not possible for speech recognizers to self-organize. It needs to be upraised in the public.

The complications involved in the speech recognition problem are well-defined by the below-mentioned features:

(i) Size of vocabulary: As the size of vocabulary increases, increases the difficulty of the task. The occurrence of verbally similar words creates a problem in recognition, like ‘WHOLE’ and ‘HOLE’.

(ii) Grammar intricacy: The complex parts of grammar also lead to the complexity of the Speech Recognition Systems.

(iii) Broken or unbroken speech: It is difficult to recognize continuous speech as compared to irregular speech. In the case of continuous speech, words are disturbed by the occurrence of co-articulation.

(iv) The number of speakers: a greater number of speakers raises the difficulty in the recognition of speech.

(v) Noise from the surroundings.

A system for speech recognition (SR), samples a speech series at a sampling frequency of 8 kHz with a precision of 8-bit, and collects a series of information at 64 kb/sec as an input. Once the execution of this series is done, a text as an output appears at a rate of about 60 bits/sec. This involves a big deduction in the size of information whereas conserving most of the useful information. A high compression ratio is required for an effective speech (above 1000:1).

To enhance the accuracy and efficiency of a recognizer it is essential to use as ample an amount of prior information as possible. It is necessary to know that there exist diverse levels of a priori knowledge. The highest level is established by prior knowledge that holds at several time instants. The bottommost level is made by prior knowledge that only holds usable information within specific contexts. In certain events of the speech recognizer, the physical attributes of the human vocal tract persist no matter the articulation, and a prior knowledge extracted from these attributes is all the time correct. On the other extreme, all the prior knowledge gathered about how a particular human says words are correct only when analysing the articulation of that human. Based on this point, speech recognizers are generally classified into two phases, as depicted in figure 1.5.

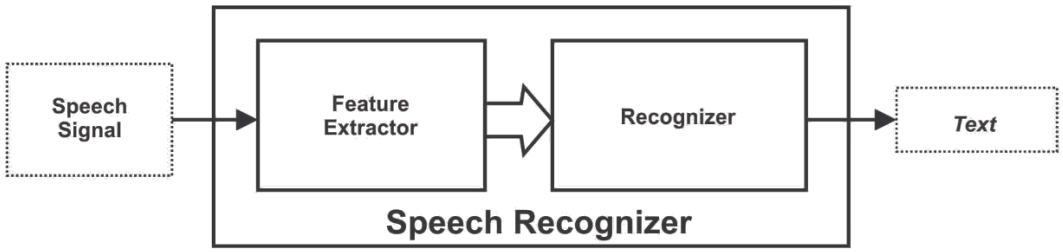


Figure 1.5: Basic building blocks of an SR [13]

The Feature Extractor (FE) block illustrated in the figure below produces a series of feature vectors, a trajectory in a certain feature space that denotes the feed in the speech signal. The Feature Extractor block utilizes the human vocal tract information to compact the data enclosed by the articulation. As it is based on a piece of true prior information, it is time-invariant. The subsequent phase is the Recognizer which carries out the trajectory

recognition and produces the exact output word. As this phase utilizes data in a particular way, a user generates articulations, adjusted by the user. The Feature Extractor block can be modelled based on the phases shown in human biology and progress. FE block converts the received sound into an internal depiction in a way it is likely to recreate the original signal from it. This phase can be modelled taking an idea from hearing organs, which firstly transduces the received air pressure waves into a fluid pressure wave and then transforms these pressure waves into neural signals. At the next stage, in the model, comes the analysation of the received data and categorizing it into the phonemes of the respective language [14]. The working of the Recognizer block is modelled based on information learned from a baby during the first six months after his birth when he familiarises his hearing organs to specifically know the vocal sound of his mother and father.

After the FE block finishes its function, the Recognizer block classifies the signal and ultimately transforms the recognized phonemes into respective words. Recognizer block behaves as if the world is made up of words and categorises every received trajectory into one word of a particular vocabulary.

It is essential to know that it is not a simple task to understand speech, it involves a very wide concept and understanding in giving meaning to the incoming information.

1.6.2 Speech Recognition Concept in Human

The whole procedure of generating and assessing speech from the thought in the speaker's brain to the formation of the speech signal, and eventually to the comprehension of the piece of information by a hearer is shown in figure 1.3. The procedure begins as soon as a thought/message appears in the brain of the talker in the upper left. The information in the message can be considered of having many kinds of depictions during the action of speech production, firstly, the information can be denoted as written content in English.

To "speak" the information, the speaker completely transforms the written content, which is the one that relates to the verbal form of the text, into an emblematic illustration of the sound series. The process of converting text symbols to the elementary sounds of a verbal form of the message using phonetic symbols is known as the language code generator (along with durational information and stress) and how the sounds are envisioned to be created, for example, the speed and emphasis.

Further actions in the procedure of speech production are the translation to "neuro-muscular controls," that is, the group of control signals that allows the neuro-muscular system to proceed with the speaker as guided, which contains the lips, tongue, jaw, teeth, and velum, in a way that is compatible with the chosen spoken message sounds and with a preferred level accentuation [15]. The neuro-muscular control actions cause articulatory motions that control the motion of vocal tract articulators in the desired way producing the preferred sounds. The final stage involved in the Speech Production method is the "vocal tract system" that generates a sound waveform that encrypts the data in the required message into the speech signal by physically creating the essential sources of sound and the suitable vocal tract movements for a certain time.

The speech perception model shows every step involved in receiving sound at the ear to interpreting the encoded message into the speech signal. The primary step includes the generation of a spectral graph from the received sound. The conversion of the sound wave into a spectral graph is carried out by a basilar membrane, which works as a non-uniform spectrum analyser within the inner ear, which spatially separates the spectral elements of the received speech signal and thus analyses them by what extent to a non-uniform filter bank. Then the further action is the neural transmission of the spectral characteristic into a sound feature set (or characteristic features discussed as in the lingual arena) that can be decrypted and executed by the brain. Next comes the transformation of the feature set of sound in the brain with the help of the language-translation process, into words, phonemes, and sentences corresponding to the received message. The final step of the speech perception model is to act or respond based on the information deduced from the previous step. In most speech perception models, one major perception of the methods is optimally primitive, however, it is usually admitted that a certain physical connection exists between the steps in the speech perception model which happens inside the human brain. Hence, the complete model is suitable to believe the procedures that happen.

The intermediate between the speech generation and perception model is the transmission channel. This transmission channel, in its simplest embodiment, involves a connection of acoustic waves between the speaker and listener residing in the same place. Along with the transmission channel there comes the noise and distortions which are essential to be considered as they exist in the real-world communication system.

1.6.3 Speech Recognition by Machine

The key objective of the SR system is to replace the hearer, though it is obvious that attaining the versatility provided by the human ear and the brain is super challenging for an artificial system. Thus, some constraints are present in SR systems. The process is dealt with in parts to upturn the performance of the recognition, and research are concentrated on those parts. The working of SR systems is contingent on the principle of roughly comparing the input data with the pre-recorded patterns [16]. These patterns are arranged and represented as words or phonemes. The data which matches best with the pattern is recognized as the symbolic representation of that data. To match raw speech signals straight is a very difficult task. A pre-processing of the signals is necessary as a significant variation of the intensity of speech signals can occur. This pre-processing is followed by Feature Extraction. Initially, feature vectors of less time period are attained from the received speech signal, and afterwards, these vectors are matched to the patterns categorized former to the comparison.

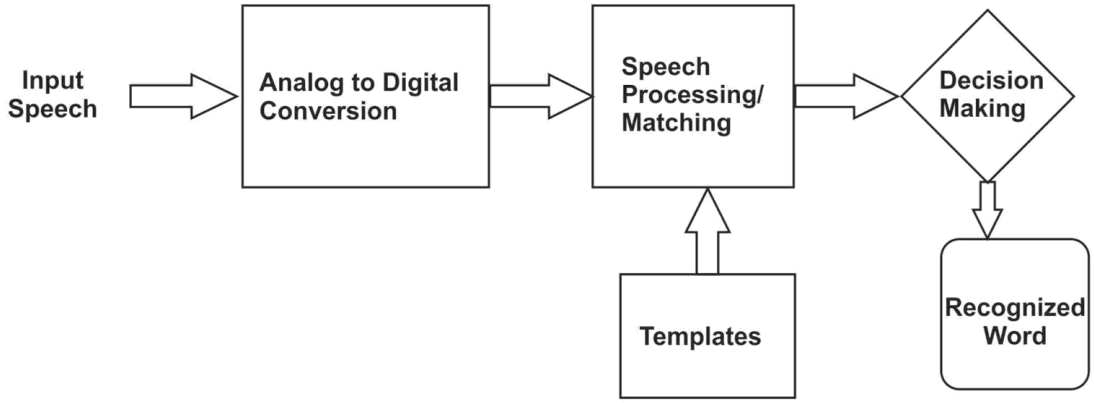


Figure 1.6: Speech recognition model in Machines[17]

Figure 1.6 shows the block diagram of the speech recognition model in machines. The first step is where the analog speech signal is converted to a digital signal. The next step is speech processing, where the digital signal is matched with the available templates. to decide output by matching the input with the template. The output block shows the recognized word output.

1.6.4 Automatic Speech Recognition (ASR)

The ASR is the method that converts the speech into digital form, classifies each sound (Phonemes) and employ a mathematical model to choose discrete word or phrase. The SR

is a special case of pattern recognition. The block diagram of the ASR application is described in figure 1.7.

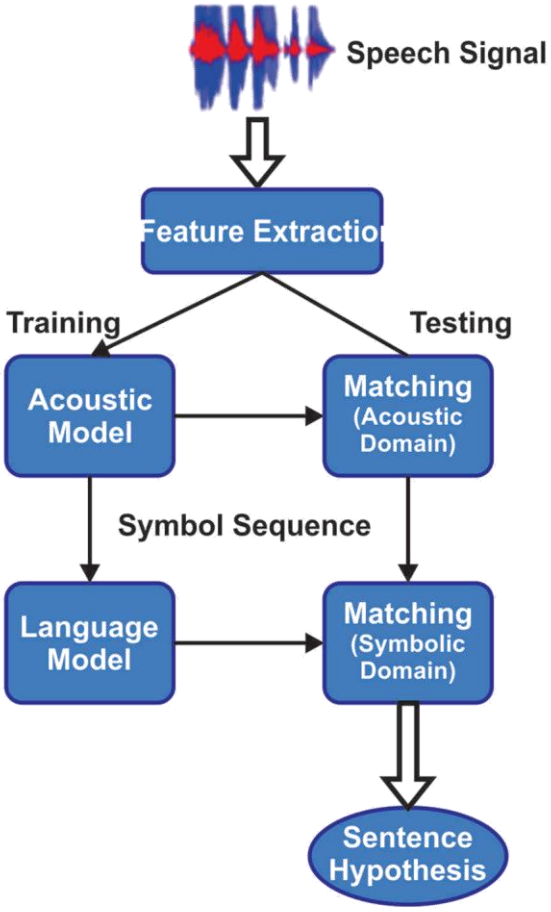


Figure 1.7: Block diagram of a typical speech recognition system

Training and Testing are the two main processes involved in supervised pattern recognition. Both training and testing work on the principle of feature extraction [18].

In the process of training, the classification model parameter is estimated utilizing a huge amount of Training data/class patterns. In the training process or recognition process, the parameter obtained from testing data is compared with the training data of every class. The testing data is confirmed to fit in the class whose model matches up a maximum with the test pattern. Finally, this technology can regulate machinery, letting freedom of functioning and higher attention to other main tasks. Simple usage of a spoken language system includes identifying sounds of the utterer, understanding them for the implementation, deducing sense from the remark, and giving a resulting response to the user (in the form of speech or some action). The information processing in the speech dialogue-based input-

output system is shown in this figure. Understanding the usage of these kinds of systems in distinct human-machine interface perspectives requires a predictive model of performance as a function of distinct related factors.

As the research in automatic speech recognition is in progress day by day, the researchers are focusing on performance improvement and real-time commercial application development.

The speaker utters the speech, from which features are extracted, matched with the trained data and classified for recognition purposes. The larger the training data, the more are the chances of better classification.

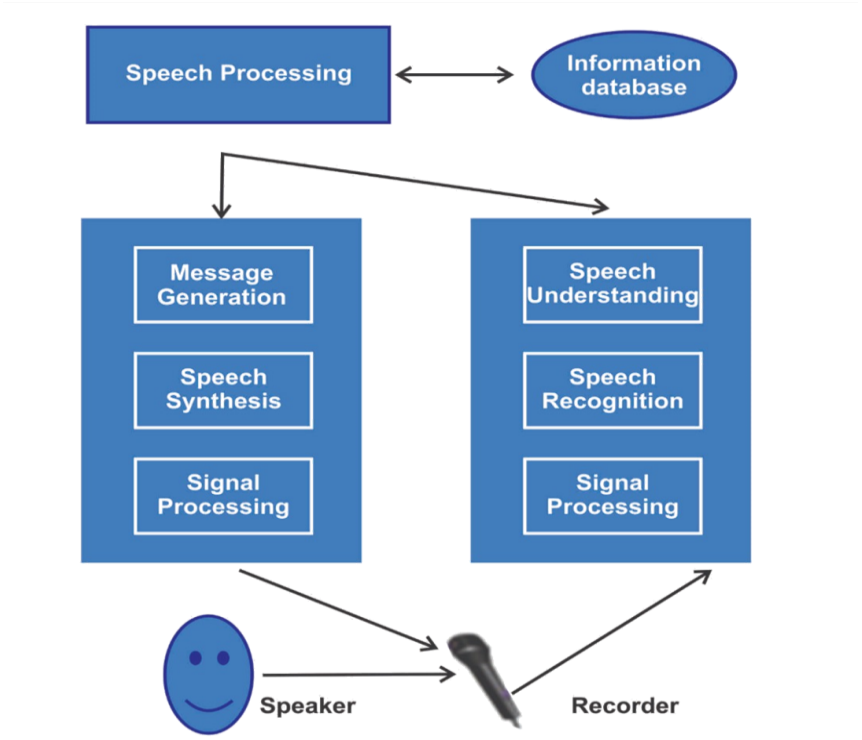


Figure 1.8: Basic architecture of the speech processing system

Figure 1.8 describes the basic architecture of a speech processing system. The information collected from the database is used for speech processing. The speakers record the speech signals, which are used for speech synthesis or message generation, etc. The processing also includes speech recognition where the speech is recognized by text.,

Figure 1.9 shows the step by step working of speech recognition systems. The training database is collected from a wide variety of speakers with different moods, ages, gender,

and class. The test data is then fed to the feature extractor for the extraction of features. The features are then matched with the available templates. The classifier then classifies the text based on available templates using certain algorithms. The recognized text is available as output.

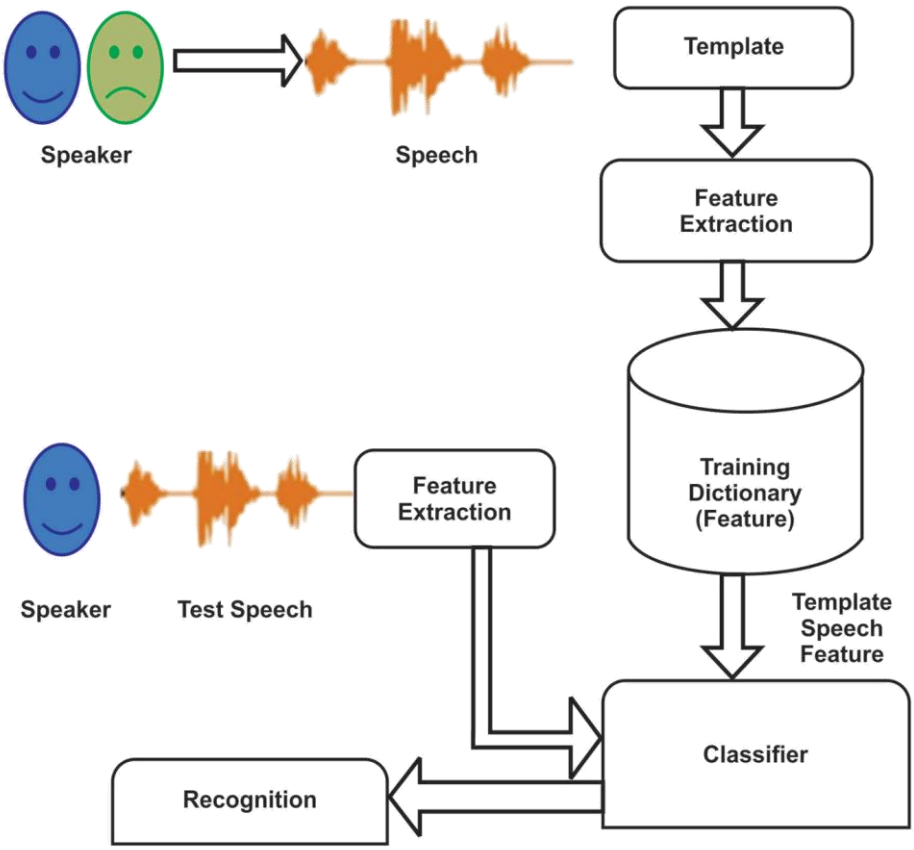


Figure 1.9: Step by step working of automatic speech recognition[19]

1.6.5 Speech Digitization

Speech processing begins with the conversion of the analog signal (air pressure being transduced to an electric signal using a microphone) to its digital form. Figure 1.10 describes the steps of analogue to digital conversion of speech. The method of converting the analogue signal to a digital signal involves two steps: sampling and quantization (Digitization).

The sampling of the signal is done by evaluating the amplitude of the signal at specific time instants. The count of samples obtained per second is defined as the sampling rate. For accurate measurements of a wave, it is required to collect a minimum of two samples in

each cycle, one from the positive portion cycle of the wave and the second one from the negative portion of the cycle. An increase in the number of samples per cycle maximizes the accuracy, but samples less than two will change the frequency of the wave. Therefore, the highest frequency wave that can be computed is one whose frequency is half the sample rate (as each cycle requires two samples). The highest frequency for a given sampling rate is known as Nyquist frequency.

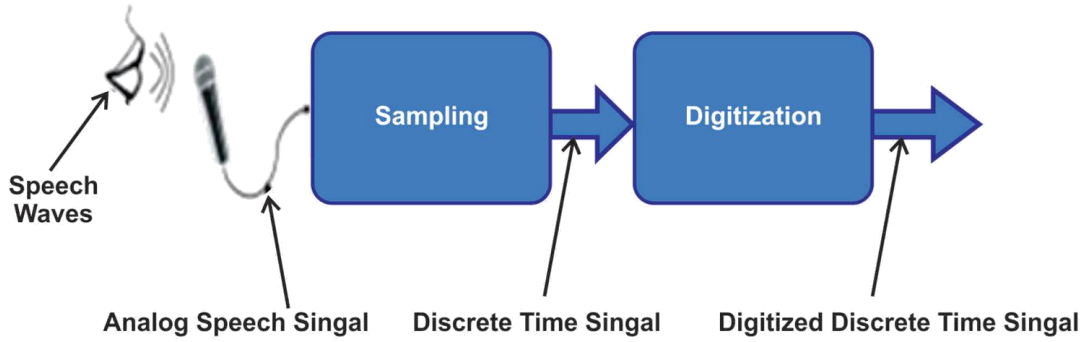


Figure 1.10: Digitization of Speech

Major information in human speech is contained in frequencies under 10,000 Hz, hence a sampling rate of 20,000 Hz would be needed to attain good accuracy [20]. But in telephonic conversations filter is applied to the speech by the switching network. The telephones are transmitted frequencies less than 4,000 Hz. The sampling rate of 8,000 Hz is adequate for a telephone-bandwidth speech such as the Switchboard corpus. The microphone speech is frequently used the 16,000 Hz (sometimes called wideband) sampling rate.

The system needs to be smart as without any clear signals from the user the system must detect pauses between the words by itself and section the speech sequence into words. This type of silence/speech indicator is known as an end-pointer, which identifies silence /speech limits (represented as dotted lines). Words can be separated by selecting speech sections between midpoints of consecutive silence sections. Maximum applications of speech utilize this kind of an end-pointer to make the user free from indicating the beginning and end of the speech.

1.6.6 Speech Recognition Process

The Speech Recognition process involves a series of processes in a particular order. These steps are enumerated in the following order. The input audio signal is passed through an acoustic processor and then matching of the words is performed to produce the text output.

(i) Input Audio: The voice of a human is transferred via a microphone attached to a computer with a normal sound card.

(ii) Acoustic Processor: The filters present in the acoustic processor eliminate the environmental noise and transform the received auditory into a series of phonemes.

(iii) Matching of words: The software attempt to match the sound of the most likely word in two ways, firstly it creates a list of probable matches that holds alike sounds, secondly, it identifies the best matching candidates using language modelling [21]. The word matching process is used for the user-defined domain.

The user can widen the domain by putting in a new word and can generate different domains for diverse applications. Ultimately, a constant speech recognizer observes appropriate info to forecast what word should come next in the current phrase. This benefits the system to differentiate among homonyms.

(iv) Decoder: Based on word matching ranking of words is done and the decoder selects the highest-ranked words.

(v) Output Text: Certain SR program includes their word processors, but several SRs let the arrangement of text straight into a distinct word processing program in an application, for example, an e-mail program or web browser.

1.6.7 Relevant Issues of Automatic Speech Recognition Design

In the current era, researchers are focusing on refining the functioning of the SR system. From the enriched literature, it is observed that the working of the system depends on the intrinsic speech variation. The performance of the system possibly is aimed to withstand a wide range of variations from the speaker.

There are main classes that affect the working of the system. Firstly, the reasonable structure of the speech signal was affected. The prominence feature of the speech depends on physiological or behavioural factors. The different physical characteristics, such as smoking habits, disease, and the environmental context make the voice softer and tense. Secondly, the long-term variation of the voice may be modified, purposefully – to transmit high-level information such as emotions. This effect is an essential part of human communication so, it is very important [22]. Third, the word pronunciation is transformed.

It may be deeply affected by accent and tone in speech recognition. To attain speaker self-sufficiency a training step is generally required and employs a definite kind of speech corpus. The relevant issues regarding the design of the ASR system are shown in table 1.1.

Table 1.1: Important problems in ASR design

ISSUES	EFFECTS
Environment	Type of noise, Signal/noise ratio, Working conditions
Transducer	Microphone, Telephone
Speakers	Speaker dependence/independence Sex, Age, physical and psychical state
Speech styles	Voice tone (quiet, normal, shouted), Production (isolated words or continuous speech read or spontaneous speech) Speed (slow, normal, fast)
Vocabulary	Characteristics of available training data, specific or generic vocabulary

(i) Speech Signal Acquisition: In speech processing, speech signal acquisition is a very significant step. The performance of the system degrades if the picked-up signal is recorded improperly. Preferably, recording requires soundproof research labs or studios. But if soundproof space is not available one should record in a silent room with very low-frequency noise. Some of the low-frequency noise sources are elevators, air-conditioner ducts, doors, mechanical systems in the buildings, heating, computer fans, 60 Hz hum from electrical appliances and water pipes. Switching off these devices during recording would help in recording a clear sound, if possible [22].

The government of India published the LDCIL standard under the department of linguistics for the recording of the speech signal. There exist so many tools like CSL, PRAAT, SPHINX, AUDACITY, and JULIUS used for recording speech databases by the researchers.

(ii) Microphone Interfaces: The microphone interface is a significant step toward the speech signal acquisition and performance of the system. The microphone is directly

connected to the computer or laptop for speech recognition. It does not show limitations instigated by the technical faults faced in the public telephone network, for example, bandwidth restrictions. In the current technological era, different companies produce microphones. The quality of speech depends on the recording environment [23]. The change in acoustic and electrical properties of the microphone lead to the system's poor performance.

The characteristic of the microphone interface which directly affects the performance of speech recognition is defined in table 1.2.

Table 1.2: The characteristic of the microphone interface

Characteristics	Example
Channel characteristics	Frequency bandwidth, distortions, echo and echo delay.
Environment characteristics	Noise type, signal-to-noise ratio.
Operating conditions	Accelerations and movements.
Microphone characteristics	Speaker distance, head-mounted or handheld.
Mechanical effects	Press pressure when talking in the microphone

The speech recognition designer must have a clue about the fact that certain users don’t find ease while interacting with a machine and don’t want to use a microphone. It is not possible to use the handheld microphone for every real-time application. For the practical SR system, the remote microphone will be of good choice [24].

The variation of the microphone during training and testing also degrades the working of the system. For the telephonic speech recognition system, the speech input is recorded using the telephone line and environment. There is a need to develop a universal adaptation microphone for a speech-based interface system that will be used for normal recording and telephone data.

(iii) Acquisition setup: To accomplish a great quality of audio the recording is carried out in a normal room in absence of noises and no echo effects. The whole recording must take place at the sampling frequency of 16000 Hz at normal room temperature and usual humidity. The speaker maintains a 12-15 cm distance facing in the direction of the microphone. A Computerized Speech Laboratory (CSL) is used for the collection of speech databases using one channel. The CSL is known as a powerful system for the analysis of voice and speech. It possesses special features which are responsible for trustworthy acoustic measurements. It is used as an input/output recording tool for a Personal Computer.

1.7 CLASSIFICATION OF SPEECH RECOGNITION SYSTEMS

1.7.1 Based on Types of Speech Utterance

An utterance is the communication (speaking) of a word or words that denote a particular sense to the computer. One-word, multiple words, a sentence, or a combination of sentences might be reflected as an utterance. SR systems are parted into some distinct classes based on the style of the speech utterance, kind of speaker, channel type and the kind of vocabulary that they have.

(i) Isolated Words: Isolated word recognizers generally need each word to have silence at starting and ending of the word. It does not indicate that it takes only a single word but needs a break amid two adjacent words. Acoustic models are constructed for each word while training the speech recognition system [25]. Hence, within the words, coarticulation is well captured by the word models. An isolated speech recognition system has three essential drawbacks. Firstly, models are defined for each word and as a result, estimation of a large parameter set is needed. Secondly, it is unnatural for speakers to speak words with pauses between the words. Hence, it is not a convenient human-computer interaction. Finally, as the vocabulary size increases, the number of word models also increases. Hence, to overcome these problems continuous speech recognition systems are used.

(ii) Connected Words: Connected word systems to some extent are alike isolated words, they let distinct utterances with the least silence in amid to be 'run-together'.

(iii) Continuous Speech: Continuous speech recognition refers to the recognition of utterances where the user can speak normally and need not pause between the words, and

this is close to how human beings speak. Continuous SR lets speakers speak naturally, even though the computer regulates the content. In other words, it's a computer transcription [26].

It takes account of a lot of "coarticulation", in which successive words combine without any apparent division or pauses in between. As continuous SR systems must use extraordinary approaches to figure out utterance limits, they are comparatively most difficult to create. Confusability between different word sequences grows as vocabulary grows larger.

Continuous speech recognition systems generally use subword sounds such as phonemes (alphabets in the case of Indian languages) as the basic unit of recognition. The sequence of phonetic models generates a bigger linguistic unit such as words/sentences while recognition. Since the number of subword units is small (typically 45 for Indian languages), it is not hard to gather enough data to train each of these basic sounds in a variety of acoustic environments. Besides, the addition of a new word is easy because only the description of a new word in terms of already trained speech sounds is to be provided.

(iv) Spontaneous speech: It is not at all a planned speech, or it can be said it is a general speech. It may include mispronunciations, false starts and non-words.

Different natural speech features for instance words being mixed or combined and even minor stammers must be controlled by an Automatic Speech Recognition system with spontaneous speech.

1.7.2 Based on the Type of Speaker Model

Every speaker has a unique special voice because of their unique physical body and personality. SR system is widely categorized into two salient classes bottomed on speaker models i.e., speaker-dependent, and speaker-independent.

(i) Speaker dependent model: This type of system is designed particularly for a specific speaker. These are usually very precise for that one speaker, but least precise for any other speaker. Though this type of system is of low cost, simpler to design and very precise but shows less flexibility.

(ii) Speaker independent model: This type of system is designed for assisting many speakers rather than a specific person. It can identify the speech patterns of a variety of

speakers. This system is very complex to design, very costly and gives relatively less accurateness. But, has the advantage of flexibility.

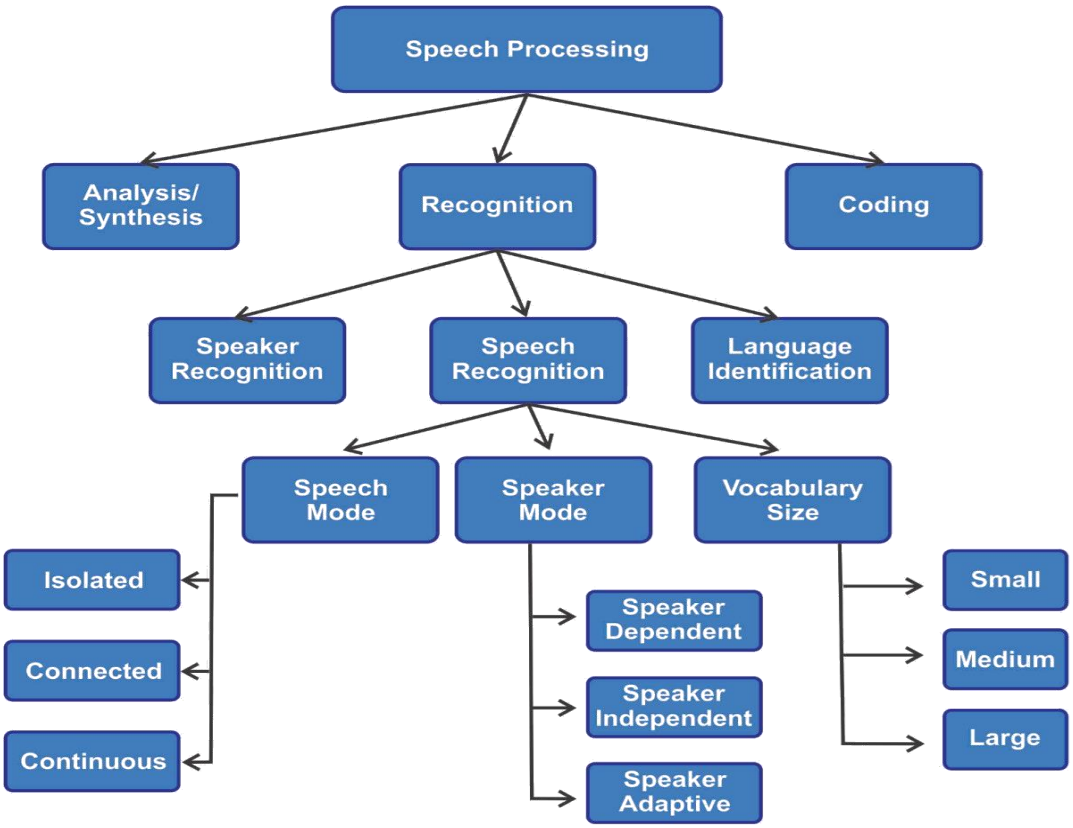


Figure 1.11: Categorization of Speech Processing[27]

1.7.3 Based on Types of Vocabulary

The difficulty, handling necessities and the correctness of the system is influenced by the vocabulary size of an SR system. However certain applications need a small number of words, while others need huge wordlists like dictation machines. There exist different categories of vocabularies in ASR systems and are mentioned underneath:

Small vocabulary – 10 to 100 words
Medium vocabulary – 100 to 1000 words
Large vocabulary – 1000 to 10,000 words
Very large vocabulary – above 10,000 words
Out-of-Vocabulary – Mapping an unknown word from the vocabulary

The variation in environs, variation in the channel, style of speaking, speaker's gender, age, and speaking speed also makes the ASR more complicated apart from the above characteristics. But the effective systems need to withstand the signal variations.

1.8 APPROACHES TO AUTOMATIC SPEECH RECOGNITION

In the last few decades, there has been so much progression in the SR field consequently improving the application of complete SR system [28]. ASR systems mainly work on three approaches 1) the Acoustic Phonetic approach 2) the Pattern Recognition approach and 3) the Artificial Intelligence approach. They are described in detail in the following sections.

1.8.1 Acoustic Phonetic Approach

In ASR systems, the features are selected based on the acoustic feature of the signal and the well-known relation between acoustic features and phonetic symbols. This is the reason why this methodology is known as an acoustic-phonetic approach. It is practically feasible, and it had been more than eight decades since researchers are working in this field.

This methodology involves the concept of acoustic phonetics that suggests that a unique and fixed number of phonetic units are present in spoken language which is further widely classified based on an evident set of properties existing in the speech signal. The acoustic characteristics of the phonetic unit are extremely unstable, both with neighbouring phonetic units and speakers. It is supposed that the protocols controlling the variance are simple and can easily be understood and implemented in real-time conditions.

The steps involved in the acoustic-phonetic approach for SR are termed the segmentation and labelling phase because it consists of the segmentation of the speech signal into discrete-time form. The schematic outline of the acoustic-phonetic approach to SR is shown in figure 1.12.

The detailed steps of the acoustic-phonetic approach are described below.

1. The analysis system of speech is the processing of the signal, which layout a proper depiction of the attributes of the time-variant speech signal. The very frequently used method for spectral analysis is the category of linear predictive coding (LPC) method and a class of filter bank techniques.

2. The feature detection stage transforms the spectral evaluation into a feature set that defines the wide-ranging acoustic characteristics of the distinct phonetic units. The feature detector has the values of the features extracted, such as pitch, energy, etc.

3. The third step involves segmentation and labelling by which the system attempts to determine static sections and then label the segmented sections by matching the region with those of individual phonetic units.

4. The consequence of the above step is the phoneme matrix from which the verbal access process finds the greatest matching word or series of words.

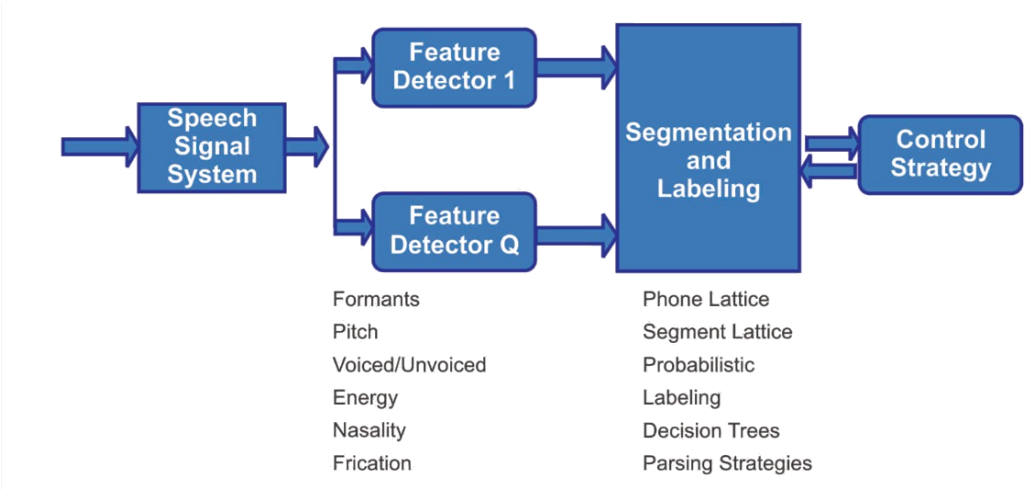


Figure 1.12: Schematic outline of the acoustic-phonetic SR system

Labelling speech and changing it into text data is a very difficult piece of work in the SR system. There exist certain tools for the labelling of natural speech. Along with dedicated tools, speech analysis software, public as well as commercial fields, also let labelling and reproduction of speech data.

1.8.2 Pattern Recognition Approach

The pattern recognition approach runs on two steps which are pattern training and pattern comparison. The important characteristic of the method is that it involves an explicit mathematical structure and forms stable speech pattern depictions.

For consistent pattern matching the set of labelled training samples using formal training, an algorithm is employed. Pattern recognition is attracting attention for the past five decades and is broadly implemented in real-world problems. A speech pattern can be

represented in terms of a statistical model or speech template, for example, HMM and it could be implemented to a word, a small sound, or a phrase.

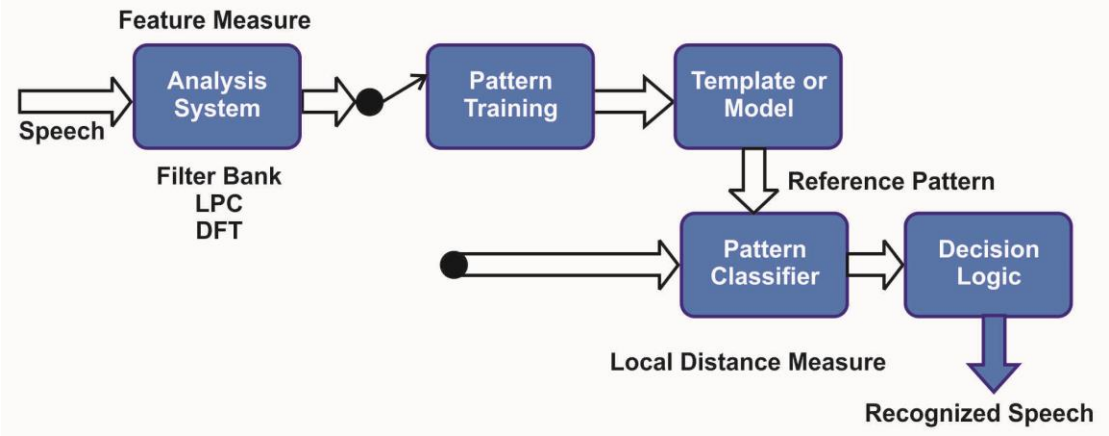


Figure 1.13: Schematic outline of pattern recognition SR

Under the pattern comparison step, a straight correlation is carried out among the unidentified speeches with every probable pattern obtained in the training phase to find the characteristic of the unidentified by the matching levels with the patterns. Over the past 60 years, the pattern-matching method has turned out the main method [29]. The schematic outline of the statistical pattern recognition method of SR is represented in figure 1.13. The pattern recognition approach has four steps.

1. Feature measurement involves the series of measurements on the feed-in signal to describe the test pattern. Feature measurement employs a spectral analysis process by using LPC and a filter bank analyser. In pattern training, single or multiple test pattern matching to speech sounds of a similar class is utilized to generate a pattern representative of the feature of the class.

2. The resultant pattern is usually termed a reference pattern.

3. In pattern classification, the unidentified test pattern is matched with all sound class reference patterns and finds out the similarity among every reference and the test pattern is enumerated. In the design logic, the reference pattern similarity count is employed to select the best matching reference pattern with the unknown test pattern.

Researchers adopted the pattern recognition approach for SR systems because of the following reasons:

(i) Ease of usage: This technique is simple to know and realize. It is reached in communication theory and calculus. It justifies a singular procedure for training and testing. It has varied usages and is understandable.

(ii) Robust and invariability: It is robust for a wide range of speech vocabularies, different users, feature sets, pattern matching algorithms, and selection protocols.

(iii) Proved great throughput: It will be made known that the pattern recognition method to SR reliably showed good throughput on any assignment that is practical for technology. It offers a clear path for outspreading the technology in a broad way such that the throughput reduces with poise as the task becomes more and more challenging.

1.8.3 Artificial Intelligence Approach

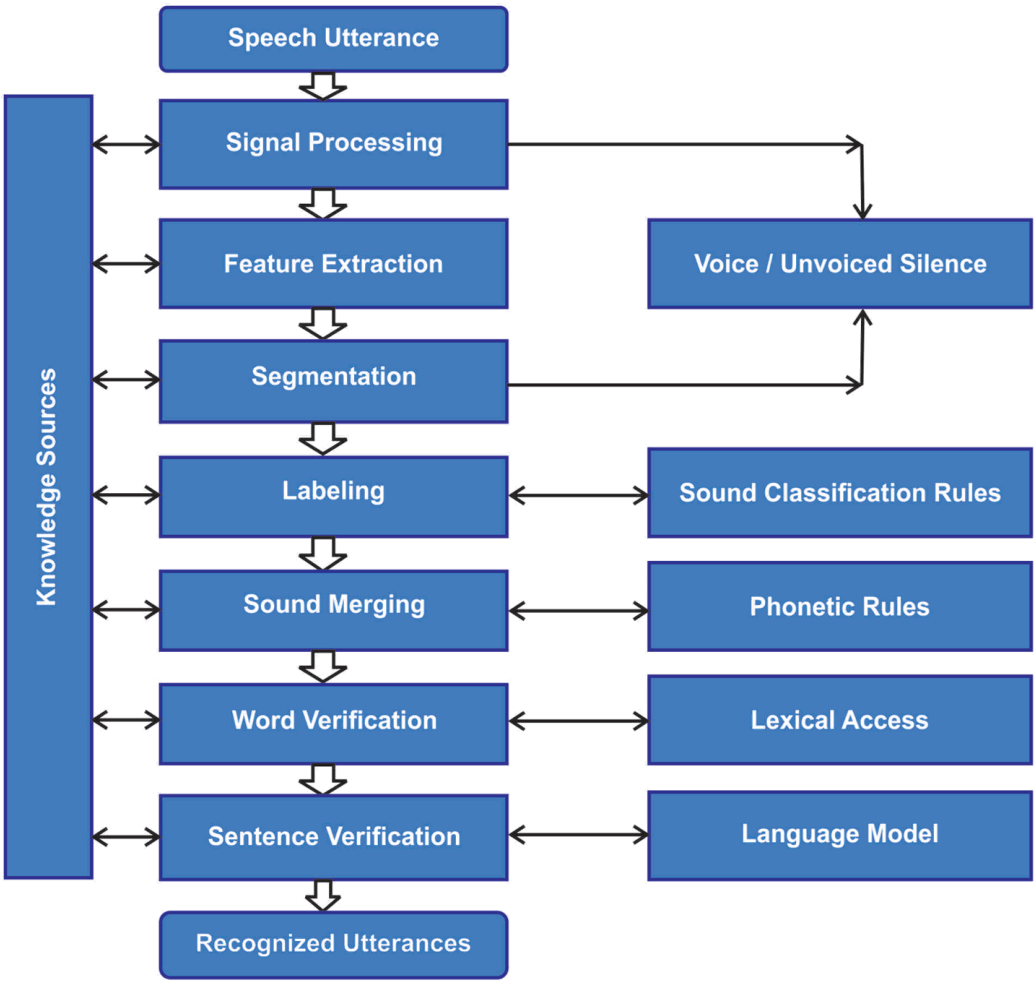


Figure 1.14: A bottom-up approach to knowledge integration of SR

The Artificial Intelligence (AI) approach to SR is a combination of the pattern recognition approach and acoustic-phonetic approach and it utilizes designs and conceptions of both.

The simple notion of the AI method to SR is to bring together the information from a wide range of sources and thus, make it a robust system to deal with any problem. Therefore, the Artificial Intelligence approach to splitting up and labelling would be to increase the usually utilized acoustic knowledge with phonetic, pragmatic, syntactic, semantic and lexical knowledge [30]. There exist so many ways to assimilate information sources within the SR. Possibly a very common method is the "bottom-up" processor, in which the lowest level processes come first higher-level processes successively to limit each phase of processing as least as possible.

A bottom-up approach to knowledge integration of Speech Recognition is shown in figure 1.14. The speech utterance is processed, its features extracted and segmented. Then they are labelled and classified. They are finally checked with phonetic rules and the language models. Thus, the final sentence is verified, and the recognised utterance is the output.

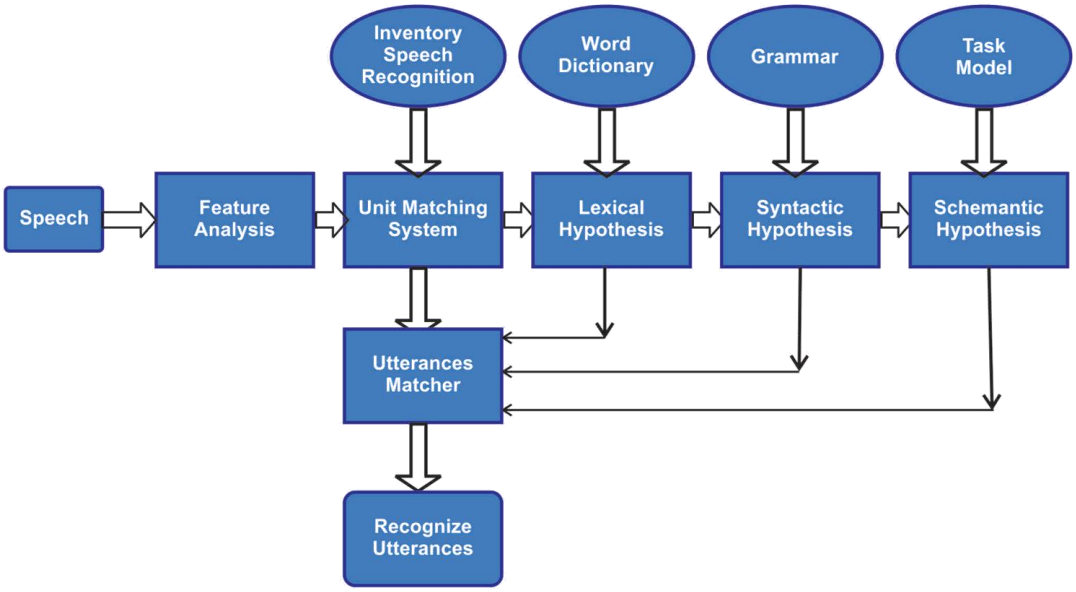


Figure 1.15: The top-down approach to knowledge integration of SR

Another way is the well-known "top-down processor", in which the language model produces word theory that is compared contrary to the speech signal and syntactically and semantically meaningful sentences are made based on word match scope. Figure 1.15 represents the system that is frequently applied in the top-down model by incorporating the unit matching, lexical decoding, and syntactic analysis module into a reliable structure.

Figure 1.15 schematically shows the top-down approach to knowledge integration of Speech Recognition. The top-down approach to the knowledge sources in speech recognition uses the speech for feature analysis. Then the unit matching system gives the output to the utterance matcher. This output is also influenced by the lexical, syntactic and semantic hypotheses. These hypotheses are in turn driven by the word dictionary, grammar and task model respectively. The final output is the recognized utterances, which is the output of the recognized matcher.

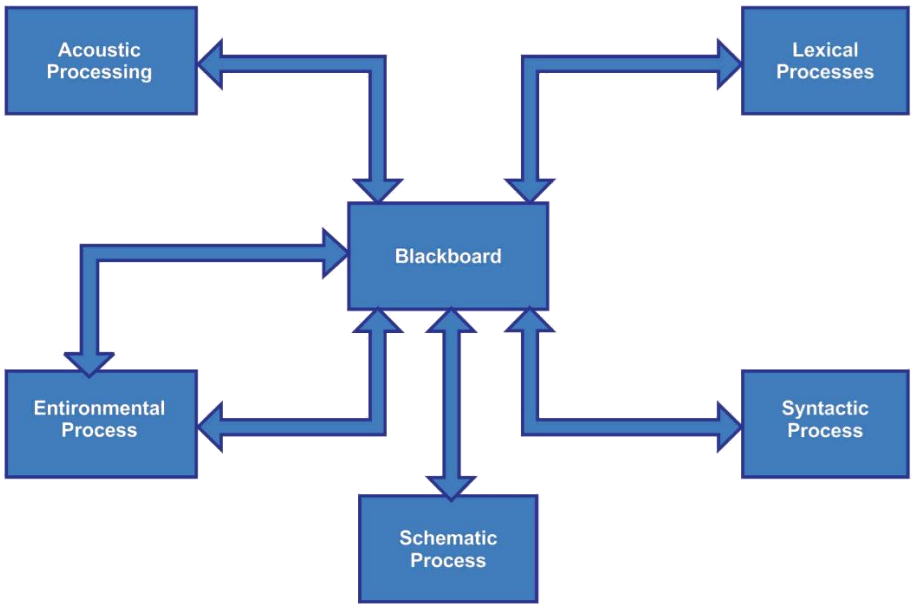


Figure 1.16: A Blackboard approach to knowledge integration of SR

One more way that exists is the "blackboard approach", which is represented in figure 1.16. Under the blackboard approach, all independent knowledge sources (KS) are considered, and assumptions and test models function as the basic means of communique between knowledge sources. Every single knowledge source is data-driven depending on occurrences of the pattern on the blackboard that match the template stated by the KS. The blackboard approach is extensively studied at CMU in 1970.

1.9 APPLICATIONS OF SPEECH RECOGNITION

ASR can play a crucial part in human-machine interactions. Computers that can identify speech in the local language may help to reap the benefit of information technology for the common man. Although any assignment that takes in interaction with a computer can use SR. Below mentioned are some of the common applications.

(i) Dictation: Nowadays it is a very popular use for ASR systems. This comprises medicinal dictations, explicit and commercial dictation, and general word processing.

(ii) Command and control: The ASR systems developed to carry out different tasks and activities on the system are termed Command and Control systems.

(iii) Telephony: Some Private Branch exchange/Voice Mail systems permit callers to say commands rather than pressing buttons.

(iv) Medical/Disabilities: A lot of people face problems in typing because of physical obstructions for example repeated pressure injuries, muscular disorder, and a lot more. For instance, people having difficulty with audibility can use a system associated with their phone to transform the caller's speech into text.

(v) Embedded applications: Certain modern cell phones contain SR that allows utterances like names. This possibly will be the most important factor in the forthcoming ASR.

1.10 CHALLENGES

Speech recognition is quite challenging due to many reasons. Some of the reasons are listed below.

Implementing a speaker-independent SR system is a challenge since the accent, dialect, and rate of speech vary across the users and geographical areas. Besides, the production of speech changes among speakers due to variations in dimensions of vocal organs due to age and gender.

If the system must recognize a large number of words that sound acoustically similar and therefore confusing, the recognition accuracy of large vocabulary ASR systems decreases.

Speech signal characteristics distort due to varying room acoustics, channel characteristics, microphone characteristics, and background noise.

The effect of co-articulation is another phenomenon that makes natural speech difficult to recognize. The physical realization of a phone can vary significantly depending on its phonetic context. The change in pronunciation happens due to word boundaries.

Spoken languages are continuous hence it is hard to find out the boundaries of each phoneme/word which constitute the sentence. Hence, it is not easy to segment and recognize speech sounds according to phoneme/word units.

1.11 CURRENT STATE OF SPEECH RECOGNITION

Speech Recognition is increasingly employed in automatic communication applications. Several research groups are involved in developing applications such as flight booking, weather forecasting, map-based navigation, and web browsing. However, the systems impose severe constraints on the users [31]. Therefore, the basic research on improved algorithms for speech recognition is actively being pursued by researchers all over the world.

1.11.1 ASR Products

Depending upon the bandwidth of the speech signal, the ability to recognize and the need for training, speech recognition systems can be Remote information access systems or Dictation systems. Remote information access systems are operated over a telephone line; hence, the bandwidth of speech signals is narrow, and the system necessarily has to be speaker independent. In 2000, researchers from MIT, Cambridge, developed an interactive flight booking system "Mercury". Mercury provided telephone gain to an online flight databank and permits users to design and price schedules amongst main airports around the globe. The system was evaluated using dialogue from users booking real flights. It is also possible to check the weather forecast by interacting with a system that gives up to date weather information over the phone.

1.11.2 ASR Research Systems

Many research groups in various languages, all over the world are researching speech recognition systems. Several toolkits have been designed for enabling research and development in continuous speech recognition such as HTK, SPHINX, and SONIC. These toolkits provide functions that can be used in speech recognition research [32]. Voice Extensible Mark-up Language is a mark-up language for generating voice user interactions that use Automatic Speech Recognition and text-to-speech synthesis systems. VoiceXML makes speech application development simpler by allowing developers to use known tools, web infrastructure and methods.

1.12 ORGANIZATION OF THESIS

The thesis has been organized in the following chapters. The dissertation will assert the topic of research. The organization of the thesis is presented in this synopsis and is as follows:

CHAPTER 1 INTRODUCTION: This chapter introduces the Automatic speech Recognition systems in detail. It describes the architectural model that the system is built on, as well as gives the breakdown of the various involved processes of speech processing, feature extraction, and classification. The chapter also includes the motivation for research, problem statement, properties of speech signals, applications, and challenges of the current ASR systems.

CHAPTER 2 LITERATURE SURVEY: This chapter gives an extensive literature review. Research papers related to the topic of speech recognition were studied. The work done by various researchers has been presented in this chapter. It also presents the history and evolution of the various ASR techniques. Finally, the chapter gives a few performance metrics for ASR systems that can be used for the performance analysis of any ASR system.

CHAPTER 3 ISOLATED WORD RECOGNITION USING ANN: This presents the Isolated Word Recognition (IWR) by using the Artificial Neural Network. This chapter introduces the problem of speech to text conversion of isolated words. The IWR problem is the first basic step for any speech recognition system because independent words are easy to understand by machines, just like humans. The chapter includes the related works of other researchers while proposing the model system for the detection of spoken words. The experimental results are included in the chapter and a discussion is presented.

CHAPTER 4 ASR USING OPTIMAL SELECTION OF FEATURES BASED ON HYBRID ABC-PSO: This chapter provides a unique hybrid algorithm of ABC-PSO for the extraction and selection of optimal features during speech recognition. The hybrid of Artificial Bee Colony and Particle Swarm Optimization is used. The experiments are conducted on the working platform of Matlab. The results are presented and discussed.

CHAPTER 5 FUZZY DWT BASED FEATURE SELECTION WITH CS-ANN CLASSIFIER: This chapter extends the study by describing a Fuzzy based Automatic Speech Recognition system. The selection of features has been done by Discrete Wavelet Transform for man-machine interaction with the CS-ANN classifier.

CHAPTER 6 CUCKOO SEARCH OPTIMIZATION BASED ANN CLASSIFIER: This chapter highlights another novel ASR by cuckoo search optimization-based ANN classifier. The novelty of the metaheuristic algorithm called CSO was suggested to train the neural network to accomplish a fast convergence rate and to upsurge the recognition accuracy.

CHAPTER 7 ASR MODEL FOR NEWS TRANSCRIPTS: This Chapter presents the work done for an ASR model on news transcripts. The collection of data (audio and text records) from a public repository to preprocessing, normalization technique, and feature extraction using MFCC and CNN classification. Finally, the experimental results are presented to show how the accuracy improved.

CHAPTER 8 CONCLUSION AND FUTURE SCOPE: This chapter discusses the future scope of the research. It also gives the contribution of the present study towards the research. And finally, the conclusion and experimental outcomes are presented. It extends the study by presenting the latest research that would be helpful to guide future researchers.

This thesis shows that the targets set as objectives are met and the outcomes achieved.

CHAPTER 2.

LITERATURE SURVEY

2.1 INTRODUCTION

Speech is the most important medium through which humans communicate. Speech has turned out to be an essential tool for Man-Machine Interaction (MMI). As in smartphones or smart devices Speech Recognition (SR) Systems are used for typing texts or initiating searches and in certain control devices it is used for on and off purposes, etc. In this chapter, the authors have tried to present a brief overview of this topic by mentioning the previous work done in this field by various researchers and progress to date. In the literature survey, they also discussed various methods and techniques used in this field to provide an ample amount of knowledge to the readers.

2.2 HISTORY OF SPEECH RECOGNITION SYSTEMS

A remarkable growth is in speech processing technology has occurred in the past eight decades. Reasons, like innovative technological interest in the process mechanism to the aspiration of things becoming automatic and building it easier using Man-Machine Interaction, attracted researchers towards the ASR technology. The evolution of this technology started in 1940 with speech evaluation and integrating systems. Homely Dudley developed the speech synthesizer in the 1930s in the Bell Telephone Laboratories and termed it as Voice Operating Demonstrator, abbreviated as VODER. The VODER was displayed in New York at the World Fair (1939) and the advanced model was verified as a significant landmark for forthcoming speaking machinery. The VODER improved by adding some electrical circuits for the synthetic creation of speech and the updated form was named the voice coder or VOCODER [7]. The VODER model was controlled by pedals and keys, but the VOCODER was a voice-controlled device. In earlier days research was based on speech integration as compared to SR but in the 1950s researchers concentrated on speech recognition systems. In 1952 Davis along with Balashek and Biddulph at Bell Laboratories developed the very first speech recognition system and named it Audrey. Isolated voice recognition in a digital format based on acoustic phonetics was observed for a single speaker and it was capable of recognizing digits only in no noise surroundings. In the 1960s isolated SR systems were established with a little vocabulary consisting of tens to a hundred words. IBM invented an SR system that may recognize sixteen words of

English speech and named it “shoebox” and was represented at the technology world fair in 1962. Though, this technique was dependent on the speaker.

In the 1980s the advancement in technology led to the development of a system that could recognize more than a thousand words. This technology showed tremendous progress in this decade. At the same time AT & T Bell Laboratories produced the self-regulating SR based speakers by using cluster-based methods and choosing self-regulating voice recognition patterns. Advancement in the technology turned into the development of new algorithms for recognition that were proficient in joining isolated words, forming connected words which led to new and advanced SR systems. Rabiner et al. [33] used Hidden Markov Models (HMM) in their work for Speech recognition. They presented an overview of the hypothesis of Markov models and presented in what manner they utilized these models for SR tasks. They implemented HMM to create a random word identifier. At present, most of the recognizers are created on the statistical basis of HMM. Since the 1980s, NNs are used for resolving classification difficulties in SR systems. With the extensive progress made in the following years, Automatic Speech Recognition expertise has progressed from speaker-dependent to no more speaker-dependent systems.

Earlier in the 1990s, some unrestricted speech models and controlled word syntax models were castoff as a format to construct vast dictionary schemes for improved learning and constant SR. The main development in this duration was the development of methods for learning speech uncertainty, the study of statistical features, and knowledge of aural and language models. The finite state transducer network was also presented along with the Finite State Machine (FSM) Library. Then at the time of application, the methods for the minimization in the size of the FSM library for the impactful implementation of vast vocabulary language understanding background were established. The FSM library is combined with a finite-state arrangement method in an integrated transducer system accompanied by the weight search. It has been a remarkable section of all new-age language recognition and empathetic background.

In the subsequent 15 years, SR systems became more powerful. They became capable of processing the infinite extent of the vocabulary. Rates of recognition also became better for practical SR tasks. Even now the ASR is seen as a typical classification difficulty. It can classify a series of words from speech waveforms. But some problems stop it from attaining the preferred and fitting results. Some of the issues include environmental noise, multi-

language recognition and multi-model recognition. Some solutions to noise removal include windowing or spectral subtraction, and Weiner filtering, and they prove to be beneficial for speech improvement. For acoustic modelling of input speech, HMM and the GMM are very suitable [34]. Machine learning technology played a crucial role in enhancing SR systems, like employing Deep Neural Networks (DNN), ANN, etc. Using these technologies has led to a spontaneous SR with unlimited vocabulary size. Authors in [35] have shown the utilization of the SVM (Support Vector Machine) technique for the recognition of emotions in ASR systems. For high dimensional vectors, SVM proves to be a low-cost classification technique. The SR systems have developed from smaller to large vocabularies, from filter banks to cepstral features, template matching to HMMs, and from speaker-dependent technology to speaker-independent technology. ANN has the advantage of taking less computation time than other systems for processing real-time speech talks, thus, they perform superiorly. Even though, when ANN have multiple hidden layers, the training process becomes time-consuming. In addition, the initialization step also impacts the ANN's performance. Hence, DNN was ideal for enormous unlabeled data [36]. Table 2.1 shows some of the exploration done in the SR field along with methods and key technologies employed for them.

in 2006 Deep learning algorithms were observed as an active field of exploration in machine learning. Machine learning algorithms improved the extraction of features and brought transformation. Several researchers stated that DNNs perform better compared to GMMs in SR systems for performing data correlation modelling and for acoustic modelling [37]. The progress in computer hardware and algorithms made end-to-end training of neural networks possible for researchers. Neural networks require less human interference and show better performance. ASR has benefited from the introduction of neural networks to a large extent.

In [38] researchers have shown transcription of audio (as input) into text for ASR systems without using intermediary phonetic representation. In [39] researchers presented an approach for automatic language identification using DNNs. In [40] researchers proposed a method, independent of language resources, to enhance the functioning of low resource ASR systems using DNN. In [41] authors presented a method for automatic feature extraction using DNNs technology, for speaker adaptive training for acoustic modelling and audio inputs, and for tracing the information flow state [42].

Table 2.1: Speech Recognition Technology over the last eight decades.

Year	Vocabulary size	Type of Speech Recognition	Methods	Key Technology
The 1960s	Small	Isolated Words	Simple Speech sounds, Phonetic properties	Analysis using Filter Banks
The 1970s	Moderate Size (100 - 1000 words)	Isolated words, Connected Digits	Template-based	Pattern recognition, Linear Predictive Coding, Zero Crossing analysis and Speech segmentation
The 1980s	Large	Connected Words	Statistical – based	HMM, Modelling for Stochastic speech
The 1990s	Large	Continuous Speech	Grammar-based Sentences.	FSM, Learning by Statistics
2000 -2005	Very Large	Spontaneous speech	Multimodal Dialog in HCI, TTS	Machine learning; ANN
2006-2010	Infinite Vocabulary size	Real-time dialogue Speech, Robust speech, Multimodal speech	Variation Bayesian estimation model, Mixed - initiative dialogue;	Deep Neural Networks (DNN), Gaussian mixture hidden Markov model (GMHMM)
2011-2015	Infinite Vocabulary size	Automatic language identification	Multimodal, Acoustic and Language modelling	DNN, GMM, DNN-HMM
2016-till date	Infinite Vocabulary size	End-to-end automatic speech recognition, On-device speech recognition	Kernel Approximation, Acoustic and Language modelling	Multi-layer Perceptron (MLP) neural network, Dynamic MLP, DNN, Kernel acoustic model

In [43] authors presented an approach for cross-language knowledge transfer using DNN.

In [44] authors presented a review on the comparison of HMM and DNN methods for speech recognition systems and tried on reducing mathematical computations required in designing SR systems for mobile devices. They evaluated their proposed method and found out that HMM performs better than dynamic MLP in SR systems, but the computation time for HMM is higher compared to DMLP. DMLP takes 9000% less time than the Hidden Markov Method for processing each pulse second of speech data. Even though, DMLP processing time increases with the increase in network size. However, this rise is less as compared to HMM system.

2.3 SPEECH RECOGNITION SYSTEMS DEVELOPED BY OTHER RESEARCHERS

The subsequent paragraphs present the various SR systems presented by other researchers. It includes their work done, methods used and inference.

Mitra et al. [45] have presented the combination of features, picking sub-band speech modulations and a summary modulation known as, the Modulation of Medium Duration Speech Amplitude (MMeDuSA) feature. They evaluated their proposed model working utilizing SRI International's DECIPHER large vocabulary continuous speech recognition (LVCSR) system, in noisy and degenerated channel Levantine Arabic speech distributed over the Defence Advanced Research Projects Agency (DARPA) Robust Automatic Speech Transcription (RATS) program. They also examined proposed models functioning in the presence of Aurora-4 noise-and-channel degraded English corpus. They have considered two different languages, Arabic and English for their work. In both languages, their proposed model showed reduced word error rates in a noisy environment in comparison to other features. They also evaluated system performance in real-world noise scenarios and channel artefacts with degraded speech. They attained consistent and good performance even in noisy and degraded channel conditions

Abdel-Hamid et al. [46] have proposed a model in which they used a hybrid NN-HMM structure for SR along with CNN. To achieve higher performance for multi-speaker SR systems they used max-pooling and local filtering in the frequency domain to stabilise speaker variance. For normalization of spectral variations, in the lowest layer of NN, they added a couple of local filtering layers and a max-pooling layer. In their experimental results, they evaluated the proposed CNN architecture in a speaker-independent SR task using the standard TIMIT data sets.

Zhizheng et al. [47] have proposed a common anti-deceiving spasm basis for the speaker verification systems, in which they use a transformed speech detector as a post-processing unit for the acceptance decision in the speaker verification system. The detector detects and claims whether the recognised speech is humanoid speech or transformed speech. A subgroup of the main task in the NIST SRE 2006 corpus was used to assess the susceptibility of the speaker verification system and the accuracy of the converted speech detector.

Khanagha et al. [48] have presented a unique methodology, titled Microcanonical Multiscale Formalism (MMF) for speech scrutiny. Their methodology centred on the accurate valuation of local scaling factors that define the inter-scale connections at every point in the signal dominion and delivers effective ways for reviewing *local* non-linear diminuendos of complex signals. Through their work, they have presented an effective manner for a guesstimate of these factors and presented that they express related info about local diminuendos of the speech signal that can be utilized for the task of splitting phonetics. They developed a two-step splitting up algorithm: in the first step they presented a new dynamic programming method to proficiently make a preliminary list of phoneme-boundary entrants and they presented hypothesis testing to upgrade the preliminary list of entrants.

Huang and Lee [49] present a new nonparametric algorithm, which utilizes the estimated system noise power and the MSE to regulate the step-size update. The inspiration of their work is that large noise in the system decreases step size and a large MSE increases step size, and vice versa. They introduced a new VSSNLMS which is simple to apply and provides better performance. They also provided theoretic performance exploration of the steady-state behaviour in their research. All-encompassing simulations depict that the steady-state behaviour forecasted by the analysis is very close to the obtained experimental results.

Amezquita-Sanchez and Adeli [50] have presented a sophisticated discussion on signal processing techniques for vibration-based SHM. They concentrated on civil structures comprising bridges and buildings. They introduced how new signal processing techniques brought up in previous years are capable candidates for upcoming SHM research. The major problem in understanding health monitoring systems in actual life is the automatic detection of loss incurred due to highly noisy data accumulated from many sensors on a

regular, weekly, and monthly basis. The new techniques for online Structural Health Monitoring must operate on noisy data efficiently, and be precise, scalable, transportable, and effective computationally.

Liu et al. [51] have discussed a new compaction technique for processing speech signals centred on the discrete cosine transform and then used the obtained densely packed signals to interfere with retrieval. They discovered a block-based big volume embedding technique, for the compressed signals embedding. In their proposed scheme, they generated a watermark by compressed signal and frame number. If watermarked speech is damaged, the damaged frames can be found by using frame numbers and recreated via compressed signal. Theoretic study and trial outcomes prove that the proposed system develops security of the speech signal by the watermark system which makes it easy to locate the damaged frames accurately and helps in recreating them.

Suman et al. [52] have proposed a new method for SR systems based on probabilistic random modelling. Speech coding methods are generally utilized in analysing and synthesis of low bit rates. Coding algorithms try to decrease the bit rate in the digital form of signal lacking an obnoxious damage quality of the signal. Various algorithms are used in speech enhancement to improvise the quality of speech. For narrowband speech signals, they used multistage vector quantization. To confirm filter stability once quantization is done, they used line spectral frequencies as a parameter for coding speech signals. In their method, they used the LBG (Linde, Buzo and Gray) algorithm for capturing the statistical information about the uncompressed signal. For quantization, they generated the codebooks using the LBG algorithm. They characterized the speech model by LPC coefficients and factorized the coefficients of the reverberation filter outputs of the multistage vector quantizer that are matched with the unconstrained vector quantization method. They evaluated results obtained from quantization regarding spectral distortion, in decibel, they measured enumeration difficulty and memory requirement in terms of Flops and Floats respectively. Results obtained from their work states that multistage vector quantization is better than unconstrained vector quantization, as well as requires less memory, even computational complexity is less and shows better performance regarding spectral distortion.

Kim and Stern [53] have presented an algorithm based on auditory processing and named it Power Normalized Cepstral Coefficients (PNCC), used for feature extraction. PNCC key

features include the usage of a power-law nonlinearity which replaced the old-style log non-linearity used in MFCC coefficients. They also proposed the usage of medium-time power analysis through which they estimated environmental factors, used for speech in addition to frequency smoothing. Their experimental results show that PNCC performs better for speech recognition than MFCC and PLP even in noisy and reverberant surroundings. The only limit PNCC show over MFCC is the slightly higher computational cost, but it can be neglected as the recognition accuracy is far better while training and testing consuming noiseless speech. Even when compared to techniques like Vector Taylor Series (VTS) and the ETSI Advanced Front End (AFE), PNCC performs better for speech recognition with very little computation.

Ghahremani et al. [54] have presented an algorithm for ASR systems that generates pitch and probability of voicing estimates, which are used as parameters for processing speech signals. These features prove to be beneficial for both tonal and non-tonal languages. Their method named the Kaldi pitch tracker is a greatly altered form of the getf0 (RAPT) algorithm. As compared to the original getf0 algorithm their method makes no tough decision whether the allotted frame is unvoiced or voiced. In their method, they assigned a pitch even to unvoiced frames while restraining the pitch trajectory to be nonstop. Their algorithm also generates a measure that can be cast off as a probability of voicing measure; it is created based on the normalized autocorrelation measure that their pitch extractor makes use of.

Li and Sim [55] have proposed a strong spectral masking system where power spectral domain masks are predicted using a DNN trained on the same filter bank features utilized for acoustic modelling. They improved their system performance by applying Linear Input Network (LIN) adaptation to both the mask estimator and the acoustic model DNNs. Subsequently, the stereo data was required for the estimation of LINs for the mask estimator, which wasn't obtainable during testing, so they suggested utilizing the LINs estimated for the acoustic model DNNs to adjust the mask estimators. Furthermore, they utilized the identical set of weights acquired from pre-training for the input layers of both the acoustic model DNNs and the mask estimator to make sure an enhanced constancy for sharing links.

Tan and Narayanan [56] have presented unique modifications of sparse group regularization methods. They stretched out the Sparse Group LASSO formulation to include diverse

learning methods for improved sparsity implementation inside a group and demonstrated the usefulness of the algorithms for spectral de-noising with applications to strong ASR systems. They showed that with a planned choice of groups better vigour to noise in SR systems can be realized when matched to methods such as the Fast Iterative Shrinkage Thresholding Algorithm (FISTA) application of the Sparse Group LASSO. Furthermore, they demonstrated that group sparse regularization methods can deliver noteworthy improvements over effective methods like the Elastic Net.

Qi et al. [57] have presented the application of the Gamma-tone filter-based feature and the experimental results on Mandarin speech data. They compared their system design with widely known PLP and MFCC's performance and concluded that their system performs significantly better even in noisy conditions. The novelty of their work is that the filter designed is purely in the time domain. That is the application of the Gamma tone filter to obtain a set of channel signals from speech signals is done in the time domain, conversely in general, in frequency domain designs signal is first converted to spectra and then filter banks are applied on them. The time-domain realization of filters has the advantage of simple hardware implementation as it uses no complex spectral computation and gives more precise results, as it avoids approximations.

Sanchis et al. [58] have proposed a feature-based and smoothed naive Bayes classification method. The proposed model is a hybrid model of generalized (word-independent) and specific (word-dependent) naive Bayes models. In statistical language modelling, the motive of the word independent model is to smooth the (class posterior) estimates delivered by the specific models. They empirically compared their classification model with confidence estimation based on posterior probabilities calculated on word graphs.

Loweimi et al. [59] have proposed a unique speech recognition system based on a phase-based feature. Their method comprises four parts: extraction using the autoregressive model (AR model), computation of Group Delay Function (GDF), compression and scale information augmentation. They coupled the AR model with GDF and obtained power spectrum estimates in a high resolution and minimum frequency leakage. The compression takes place in two steps as in MFCC without calculating a logarithm of the output energies. The last step uses the Hilbert transform relations and accompaniments the phase spectrum info by augmenting the phase-based feature vector with scale information.

Zhou et al. [60] have proposed an innovative acoustic modelling structure in which they used a single DNN to calculate the posterior probabilities of tied HMM instead of using multiple DNNs. In their proposed method, they grouped all tied states of context-dependent HMMs into some disjoined clusters contingent on the training data related to these HMM states. After this, using multiple GPUs they separately trained some hierarchically structured DNNs for these disjoined clusters of data. At the time of decoding, they calculated the final posterior probability of each tied HMM contingent on output posteriors from multiple DNNs.

Delcroix et al. [61] have introduced an SR scheme that can identify speech in the existence of many noise generators changing with time, for example, noises originating in a usual group of persons in a room. They used features like spectral, spatial features (directional) and temporal features in their proposed model to differentiate between noise and speech when operating in a severe noisy environment. This is cognized with a model-based speech improving pre-processor comprising of two matching features, for spectral and spatial information they used a multi-channel speech–noise separation method, succeeded by a single channel enhancement algorithm that employs the prolonged time-based features of the speech acquired from examples of clear speech. Furthermore, they employed an adaptation technique to recoup for any incompatibility that may exist between the acoustic model and enhanced speech, which merges the vigorous adaptable advantage of the variance of the Gaussians of the acoustic model with conventional maximum likelihood linear regression.

Xue et al. [62] have proposed a hybrid SR system, a combination of Neural networks and HMM techniques, for speaker adaptation based on Singular Value Decomposition (SVD). They applied SVD on the weight matrices in trained DNNs and then adjusted rectangular diagonal matrices with the flexible data. Updating the weight matrices by altering the singular values removes the over-fitting problem. They tested their proposed SR model using two standard SR tasks, large vocabulary speech recognition and TIMIT phone recognition in the Switchboard task.

Deng et al. [63] have presented a method to overcome the issue of parallelizing learning algorithms for deep architectures by employing Deep Stacking Network (DSN). The DSN lay out a process of stacking easy executable modules in constructing deep architectures, with a convex learning problem in each module. Further calibrating improved the Deep

Stacking Network, even though introduced slight non-convexity. In the Deep Stacking Network, full learning is in batch-mode, building it responsive to parallel training over several machines and hence possibly scalable over the large training data.

Keronen et al. [64] have presented an ASR system using a missing data approach to recompense severe noisy conditions comprising both convoluted and additive components. They used different variety of acoustic features in the Gaussian mixture model (GMM) classifier to identify the missing (noise corrupted) and the unpredictable components. To do SR with the partly perceived data, they substituted the missing components with clear speech estimates calculated using both cluster based GMM imputation and sparse imputation. Their proposed missing data approach when assessed on the CHiME reverberant multisource environment corpus, shows improved keyword accuracy rates significantly in all SNR scenarios when matched to two related mask estimation methods formed on time difference pairs and inter-aural level. When compared, it is found that cluster-based imputation performs better than sparse imputation. The system achieved the highest keyword accuracy and became immune to imputation errors when trained on imputed data.

Yeh et al. [65] have proposed a novel new structure for the identification of greatly imbalanced code-mixed bi-lingual speech using an extra frame-level language identifier in the traditional SR system. They also proposed the use of Blurred posteriorgram features (BPFs) in the language detector. They evaluated their method using real and unprompted lectures delivered by National Taiwan University. The task of speech recognition becomes difficult in greatly imbalanced language sharing in code-mixed speech. The result shows that the system performs better due to bigger training data and an enhanced adaptation approach.

Sun and Lee [66] have proposed unique methods for balancing the modulation spectrum for the extraction of features in SR systems. Step of the transformation of the temporal trajectory of the extracted features into the magnitude modulation spectrum is common in all methods. In spectral histogram equalization (SHE) and two-band spectral histogram equalization (2B-SHE), they equalized the histogram of the modulation spectrum for every expression to a reference histogram attained from noiseless training samples or performed the balancing with two sub-bands on the modulation spectrum. In magnitude ratio equalization (MRE), they defined the amplitude proportion of minor to major modulation

frequency elements for every expression and equalized that to a benchmark value attained from clean training d noiseless training samples. Proposed methods can be regarded as time-based filters that are reformed for every testing expression.

Smaragdis and Raj [67] have proposed a new Markov model to recognize speech. Their system can recognize speech from footage of concurrently speaking a prior identified speaker. Their work relies on the latest work on the positive description of spectrograms, which is greatly useful in problems of source separation. In their approach, their Markov selection model is capable of recognizing series even if present in a mixed form that is there is no requirement to perform the signal separation. If compared to the factorial Markov model, the proposed approach shows reduced complexity in computation and state-space with a linear number of sources.

Muhammad et al. [68] have proposed a feature-based ASR system that works with a reduced number of computations yet gives high accuracy. The feature they used is the spectro-temporal directional derivative (STDD) feature. The suggested STDD feature is obtained by putting distinct directional derivative filters in the spectro-temporal domain. By using discrete cosine transform they compressed the feature dimension. The tests were done with Arabic numerals vocal sound samples vocalised by individuals with and without voice pathology.

Rennie et al. [69] have presented a method for speech recognition using the Factorial Hidden Restricted Boltzmann Machine (FHRBM). Noise and Speech are demonstrated as autonomous RBMs, and the interface among them clearly showed how noise and speech merged to produce detected noise interrupted speech features. By comparison with RBMs, where the bottommost layer of random variables is detected, inference in the FHRBM is uncontrollable, ascending with a rise in hidden units. For an effectively estimated inference, they introduced variation algorithms that ascend linearly with an increase in hidden units. FHRBM when matched with customary GMMs based on noisy speech factorial models, showed an advantage that both noise and speech are highly distributed, letting the model learn a parts-based representation of noise interrupted speech signal that can simplify well to formerly hidden noise configurations. Obtained outcomes are favourable.

Alam et al. [70] have presented a method for robust continuous speech recognition based on cepstral features with normalized Minimum Variance Distortion-less Response

(MVDR). The widely used mel-frequency cepstral coefficient (MFCC) features for speech recognition are calculated from a direct spectrum estimate, that is, by taking the DFT of the speech frames and squaring the magnitude. Non-parametric estimators execute below par under contrary and noisy conditions. To increase the robustness of the SR system they proposed more robust features based on the normalized MVDR method.

Ting et al. [71] have presented a unique Self-Adjustable Neural Network (SANN), to facilitate the network to adjust the conferring to distinct data input sizes. The suggested method is applied to the SR of TIMIT isolated words and Malay vowels. SANN is the benchmark against the Hidden Markov Model (HMM).

Kashiwagi et al. [72] have proposed the utilization of DNNs to increase traditional methods of statistical feature improvement centred on piecewise linear transformation. Stereo-based piecewise linear compensation for environments (SPLICE) is used to model the probabilistic distribution of input noisy features as a mixture of Gaussians. Particularly in linear conversion, the system performance degrades. To model a non-linear relationship among clean and noisy feature spaces they used feature enhancement as a tool based on neural networks technology. But it likely experiences over-fitting problems. The authors attempted to alleviate this problem by minimizing the total model parameters used for estimation. They trained their neural network output layer with the states in the clean feature space, other than doing it in a noisy feature space. Their approach made the extent of the output layer independent of any type of noisy environment.

Dighe et al. [73] have proposed a method to model the acoustic space of DNN class-conditional posterior probabilities as a combination of low-dimensional subspaces. For this purpose, the training posteriors are used for sparse coding and dictionary learning. Sparse depiction of the test posteriors with this dictionary assists projection of the space of training data. Depending on the datum that the intrinsic dimensions of the posterior subspaces are very insignificant and the matrix of all posteriors fitting to a class has a very low rank, they exhibited in what manner low-dimensional structures allow additional improvement of the posteriors and correct the false inaccuracies because of mismatch situation.

Li and Fung [74] have proposed a combined structure for huge vocabulary continuous variegated language SR that controls the pronunciation outcome in the bi-linguistic acoustic model and the inversion constriction familiar to the interpreter in the language

model. Their asymmetric acoustic model with telephone set extension advances upon earlier work by imposing stability among phonetic knowledge and data. Their proposed language model advances upon earlier work by (1) predicting code-switching points in the mixed model using the inversion constraint and (2) mixing a translation model, code-switch prediction model, and a reconstruction model. This combination infers that their language model evades the drawback of communicated error that might ascend from dissociating these phases. To conclude, a WFST-based decoder mixes the acoustic models, the code-switch language model and a monolingual language model in the matrix language altogether.

Xiao et al. [75] have proposed a method for speech recognition based on speech features, which is a framework for the combined regularisation of spectral and temporal statistics. Existing regularisation approaches for features, regularize the temporal and spectral characteristics of feature statistics distinctly to deal with echo and noise. Consequently, the interface between the temporal normalization e.g., Temporal Structure Normalization (TSN) and the spectral normalization e.g., Mean and Variance normalization (MVN) is overlooked. They proposed a joint spectral and temporal normalization (JSTN) structure to instantaneously regularize the two characteristics of feature statistics. In JSTN they filtered the feature trajectories by linear filters and optimized the filter coefficients by making the most of a likelihood-based independent function.

Zhang et al. [76] have used a set of SVM kernel functions i.e., an Optimal Relaxation Factor (ORF) kernel function for SR and demonstrate that the ORF function is a Mercer kernel function. The testing shows that the ORF kernel function is effective in mapping trends, bi-spiral, and issues in SR. Their research concludes that the ORF kernel function does well than the Kernel with Moderate Decreasing (KMOD), the Exponential Radial Basis Function (ERBF) and the Radial Basis Function (RBF). Moreover, the outcomes of SR with the ORF kernel function show greater accurateness in recognizing.

Dehzangi et al. [77] have proposed a unique technique to get a selective feature conversion depending on the output coding technique for ASR. The speech features were projected to a new space from their original space using the output coding transformation. In the new feature space every facet of the features size statistics to differentiate between different phones. They expanded the small duration spectral features using polynomial expansion to a high-dimensional space where the general linear discriminant sequence kernel is

implemented to the series of input feature vectors. Then, formulating the output coding conversion through a set of linear SVMs projects the series of high dimensional vectors into a controllable low-dimensional feature space where the consequential features are isolated continuous output codes for the successive multi-class classification task.

Imseng et al. [78] have proposed a unique method based on Kullback-Leibler (KL) divergence that can make use of multi-lingual info in the form of widespread phoneme posterior probabilities trained on the acoustics. They expressed a way to coach a recognizer on some different languages, and afterwards identify speech in a target language for which only a little quantity of information is obtainable. They did a test using Greek SpeechDat (II) data and showed that their proposed design is sound and showed that it can perform better than a current modern HMM/GMM system. They also used a hybrid Tandem-like system to understand the extra benefit from the source.

Zhang et al. [79] have proposed PAC-RNN i.e., the prediction-adaptation-correction RNN, in which estimated the state posterior probability using a correction DNN depending on both the current frame and the prediction formed on the past frames by a prediction DNN. The outputs obtained from the central DNN are feedback to the prediction DNN to produce finer estimates for the future frames. In the presented model using the current frame info and the feedback info, the central DNN modifies the estimate given by the prediction DNN. On the other hand, it is visible that adjusting the main DNN's behaviour relies on the prediction of the DNN.

Zouhir and Ouni [80] have proposed an ASR method based on feature extraction in noisy situations. Their work is inspired by the human hearing system that acts out the middle/ outer ear filtering by a low-pass filter and the spectral behaviour of the cochlea by the Gammachirp auditory filterbank (GcFB). The proposed model is tested on speech samples corrupted by real-life noises. The obtained results suggest that the proposed model performs better than classic methods like Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), Mel Frequency Cepstral Coefficients (MFCC) and Linear Prediction Cepstral coefficients (LPCC). The proposed model for the ASR system is based on HMM with continuous Gaussian Mixture densities (HMM-GM).

Nugraha et al. [81] have presented a method based on NNs for feature enhancement to plot the reverberant feature in a log-mel spectral dominion to its conforming anechoic feature.

They used the Cascade2 algorithm for mapping which is implemented by cascading the trained NNs and segmentation-based regularization. They carried out experiments using ASR and speaker identification (SID) systems to assess the proposed method. They used CENSREC-4 datasets for ASR systems and for the SID system they used their own simulated and real reflected datasets. It is concluded from the results that by using low speaker variant data as training data and limited stereo data, the suggested approach could considerably advance the functioning of both ASR and SID systems.

Zhao et al. [82] have presented a strong SR system utilizing a microphone array for the 3rd CHiME Experiment. A minimum variance distortionless response (MVDR) beamformer with adaptive microphone gains is suggested for vigorous beamforming. Two microphone gain estimation methods are considered with the speech-dominant time-frequency bins. A multichannel noise reduction (MCNR) post-processing is suggested as well to decrease the intervention in the MVDR treated signal.

Esfandian et al. [83] have proposed a method for secondary feature extraction and selection based on clustering in spectro-temporal dominion. To decrease the dimension of the feature space in the spectro-temporal domain, in the suggested approach they applied two clustering techniques, namely, weighted *K*-means (WKM) and Gaussian mixture models (GMM). The components of centroid vectors and covariance matrices of clusters are measured as features of the secondary feature vector of each frame. To assess the effectiveness of the suggested approach, they carried out tests for new feature vectors on the classification of phonemes in the main types of phonemes in the TIMIT database. It was revealed that by using the suggested secondary feature vector, a noteworthy enhancement was shown in the rate of classification of different sets of phonemes relating to MFCC features.

Kurakin et al. [84] have proposed a practical system for the recognition of dynamic hand gestures. The proposed model is completely robust and automatic to variations in style and speed along with hand orientations. Their technique is based on an action graph, which shares alike robust properties with standard HMM but needs less training data by permitting states to share among different gestures. They developed a new technique to deal with hand orientations, for hand segmentation and orientation normalization. The suggested system is assessed on a perplexing dataset of twelve dynamic American Sign Language (ASL) gestures.

2.4 PERFORMANCE EVALUATION OF ASR SYSTEMS

When a new system is developed, its performance must be evaluated. This is necessary to check the validity of results and to compare the new system with other systems developed and in use.

Accuracy and speed are considered parameters to check the speech recognition system. The accuracy of the system is identified by the percentage of the recognized words whereas, speed relies on the computation time. The various check parameters for speech recognition systems are:

Word Recognition Rate (WRR) represents the a*ccuracy* of the system, equation 2.1 is used to find WRR as follows:

$$WRR = H - 1/N \tag{2.1}$$

The set of Words recognized correctly is denoted by H and the set of words is denoted by N.

The system performance is measured using ***Word Error Rate*** using equation 2.2,

$$WER = 1 - WRR \tag{2.2}$$

Sensitivity and Specificity are the terms used to recognize the test samples. The presence of positive test results gives the ***False Positive Rate (FPR)*** and is found by equation 2.3

$$FPR = FP/(FP + TN) \tag{2.3}$$

The capability of a technique to identify the negative samples correctly is defined as ***Specificity.*** It is given by equation 2.4

$$Specificity = TN/(TN + FP) \tag{2.4}$$

The capability of a technique to identify the positive samples properly is defined as *Sensitivity*. It is given by equation 2.5

$$Sensitivity = TP/(TP + FN) \tag{2.5}$$

Where TP stands for True Positive, TN stands for True Negative, FP stands for False Positive, and FN stands for False Negative.

Expansion of common input features together with ANN delivers accuracy improvement in the ASR systems. Accuracy is a very crucial parameter to measure the performance of the SR system.

2.5 RELATED WORK TO THE PROPOSED TECHNOLOGIES

The related work in the field of speech intelligibility is mentioned underneath:

The three emotional conditions: Cheerful, neutral and unhappy recognized by Yixiong et al. Mel-energy range powerful coefficients (MEDC), vitality message, MFCC and linear predictive range HTML coding (LPCC) are the researched characteristics. For Training, the Support Vector Machine (SVM) classifier uses the self-built Chinese emotional databases and German Corpus (Berlin Database of Emotional Speech). The resultant section compares the various combinations of features in unlike databases. The experiment shows that on the combination of energy+MEDC+MFCC the result revealed shows the highest degree of accuracy on both the Berlin emotional database and Chinese emotional database.

A system was proposed by Kaur and Singh [86] to identify the speech which depends upon the best probable element together with the dialect. Identifying an individual from a spoken phrase uses speaker recognition. Neural networks (NN) provide the basis to develop a speech recognition system. They developed the SR system based NNs and used 23 phrases for speech recognition while utilizing the obtained features REMOTE CONTROL, LPC and LPCC from talk transmission along with created element vectors. For training and recognition procedures of several languages and speakers, NN back-propagation learning algorithm is used. The database consists of twenty-five speakers. Their ANN model contains five hundred seventy-five neurons in the input layer and twenty-five neurons in the outcome layer and from the obtained experimental outcomes it is noticed that the average identification accuracy achieved is 93.38%.

Poonkuzhali et al. [85] have proposed a method for ASR and feature selection algorithm using Ant Colony Optimization (ACO). They considered the original speech signal as input and employed the MFCC feature extraction method and extracted thirty-nine coefficients. The Ant Colony Optimization (ACO) technique appeared helpful to withdraw the unwanted features. The new validation with their approach showed that the last number linked with features detached knowingly.

Petkov et al. [86] had suggested a speech pre-enhancement technique centred on the logic of matching the original message with the recognized. This qualifying criterion was properly approximated because the likelihood with a suitable transcript provides a decent approximation of the noisy features of speech. In the profile of atmospheric noise, along with a decline in the actual signal-to-noise relation, speech smartly drops off. They implemented a pre-enhancement process on speech that improves the suggested fitting parameter to a couple of variables of distinct speech reform schemes enclosed by a better energy conservation constraint. The suggested procedure was necessary for former information of transcription with the transmitted message and acoustic speech models from a computer-based SR system. The proposed scheme showed extensive growth showing clear speech results and provided a baseline system that improves the perceptual distortion-based unambiguity degree.

Joshi and Cheeran [87] have presented an automatic speech recognition technique using a back-propagation NN (BP-NN). The approaches adopted in this research may be extended to certain general applications which include missile tracking, identifying targets in the sonar system, and organizing underwater acoustic signals. The BP-NN algorithm worked on the offered training samples and their perfect result beliefs to observe patterns through altering the actual service theories related to the nodules and weights of various connections linking its nodules. An actual proficient interconnection is perceived far along with practices recognition of features within the Automatic Speech Recognition platform.

Baskent et al. [88] have proposed a review on the improvement of speech for persons with impaired ability to hear using phonemic restoration (PR). There are two conditions used to measure the speech signal interrupted occasionally and its realization is disturbances free, and interruptions loaded with loud commotion. Linear amplification was used while numerous modern listening devices provided complete intensification. The purpose of this selection was that the study was initiated to provide standard phonemic restoration data with impaired hearing listeners, and extensive resolutions could have had unexpected effects due to their nonlinear character.

Liu et al. [89] have investigated cross adaptation of language model, either execute as an autonomous hybrid approach or employed in conjunction with auditory model cross tuning to enhance huge vocabulary continuous speech recognition (LVCSR) method. Language models are of three kinds which contain a multi-level LM that models each syllable along

with arrangements of the word, a word-level NN LM, and these two were cross adjusted in the linear combination. The experimental findings on a contemporary SR assignment recommended complementary characteristics that occur on manifold layers of the advanced grading between extremely different sub-frameworks.

Kim and Stern [90] had proposed three methods developed using missing-feature methods to improve the accuracy of the SR system. The frequency-dependent classification was the first innovation of this research, which engross independent classification. Coloured-noise creation using multi-band partitioning was the second innovation, which included the usage of hiding noises with feigned hosted temporal and spectral variation in training the Bayesian classifier. A flexible technique to evaluate the inferred estimates of the mask classifier was the third innovation, which ascertained if an exact frequency-time segment of the experimental data was invalid or else valid. It determines that suggested progressions offer enhanced accuracy in SR on a compact size vocabulary trial.

Zhang and Gales [91] have described a specific model in this system, structured support vector machines (SSVM), and how it can be applied to medium to large vocabulary speech recognition tasks. An essential aspect of SSVMs was the form of the joint feature spaces. Here, context-dependent generative models, hidden Markov models, are used to obtain the features. First, the features extracted are a function of the segmentation of the utterance. A Viterbi-like system for obtaining the "optimal" segmentation was described. Second, SSVMs can be viewed as large margin log-linear models using a zero-mean Gaussian before the discriminative parameter. However, this form of prior was not appropriate for all features. A modified training algorithm was proposed that allowed general Gaussian priors to be incorporated into the large margin criterion. Finally, to speed up the training process, a 1-slack algorithm, caching competing hypotheses and parallelization strategies are also described.

Wu et al. [92] have extended the previous work by exploring a more efficient convex optimization method with the technique of second-order cone programming (SOCP). More specifically, they have studied and proposed several SOCP relaxation techniques to convert LME of HMMs in speech recognition into a standard SOCP problem so that LME can be solved with more efficient SOCP methods. The proposed LME/SOCP approaches have been evaluated on two standard speech recognition tasks. The experimental results on the TIDIGITS task showed that the SOCP method significantly outperforms the gradient

descent method and achieves comparable performance with SDP, but with 20-200 times faster speed, requiring less memory and computing resources.

Hirayama et al. [93] have developed a method of recognizing mixed dialect utterances with multiple dialect language models by using small parallel corpora of the common language and a dialect and a large common language linguistic corpus. Their two main methods were maximization of recognition likelihoods and integration of recognition results. The former constructed mixed dialect language models by using the weighted average of n-gram probabilities and pronunciation probabilities of language models to be mixed. The mixing proportion was chosen for each utterance so that it achieved the largest recognition likelihood of the utterance. One of the main contributions of this paper was that the estimates were conducted for individual utterances and not for individual speakers. The latter integrated recognition results with each single language dialect model to output results that included fewer errors. The former output results in higher recognition accuracies, while the latter improved results with lower calculation costs.

Gharavian et al. [94] have investigated the effect of MFCCs, energy, formant- and pitch-related features on improving the performance of emotion recognition systems. To compensate for the effect of emotion on the recognition rate, the normalized values of formants have been used as supplementary features. The normalization has been performed using a DTW-MLP hybrid model after determining the frequency ranges that are most affected by emotion. To decrease the computational load, FCBF and ANOVA feature selection methods have been employed. In this way, various combinations of the features have been selected by feature selection algorithms. The performance of the proposed system has been compared with some other emotion recognition systems.

Haque et al. [93] have proposed a method of unequal (asymmetric) compression, i.e., higher compression applied in the higher frequency regions than the lower frequency regions. The methods were applied and tested in the MFCC and the PLP parameterizations in the spectral domain, and the ZCPA auditory model was used as an ASR front-end in the temporal domain. The extent of the asymmetric compression was applied as a multiplicative gain to the existing static compression and was determined from the gradient of the piece-wise linear segment of the perceptual compression curve. The proposed method has the advantage of adjusting compression parametrically for improved ASR performance and audibility in noise conditions by low-frequency spectral enhancement, particularly of

vowels with lower formants. Continuous density HMM recognition using the Aurora 2 corpus and the TIdigits showed performance improvement in additive noise conditions.

Kaya et al. [94] have extended a new preferential projection centred feature selection approach by employing the power of stochasticity to deal with local minima and to minimize the computation complication. The method assigned weights both to groups and to features independently in various arbitrarily chosen contexts and then combined them for an ultimate ranking. The efficiency of the approach was presented in the latest communicative challenge corpus to identify the degree of clash in binary and group chats. They advanced the contemporary in this corpus using the INTERSPEECH 2013 Challenge protocol.

Cumani and Laface [97] have validated that a very small subset of the training pairs was essential to train the original PSVM model and proposed two methods that permitted removing most of the training pairs that were not necessary, without damaging the precision of the model. This permitted intensely decreasing the computational resources and memory necessary for training, which became possible with enormous datasets comprising many speakers. They had evaluated these methods on the extended core situations of the NIST 2012 Speaker Recognition Estimation. Their consequences displayed that the precision of the PSVM trained with an appropriate number of speakers was 10%-30% better associated with the one attained by a PLDA model, reliant on the testing conditions. Since the PSVM precision expanded with the training set size, but for large numbers of speakers PSVM training did not scale well, their selection methods became applicable for accurate training discriminative classifiers.

To optimize a feature vector, Chatterjee and Kleijn et al. [98] had developed a new framework such that it imitates the human auditory system behaviour. In an offline manner, the optimization was conceded based on the assumption that the local geometries of the feature vector domain and the perceptual auditory domain must be analogous. Accompanied by a static spectral auditory model, using this principle they optimized and modified the static spectral Mel frequency cepstral coefficients (MFCCs) with no feedback from the SR system. Formerly the task was prolonged to comprise spectro-temporal auditory characteristics into manipulating a new dynamic spectro-temporal feature vector. Making use of a spectro-temporal acoustic model, the dynamic feature vector was planned and improved to integrate the human hearing response behaviour across frequency and

time. They exhibited that an essential development in ASR functioning was achieved for whichever background situation, clean in addition to noise.

Dikici et al. [99] have presented certain ways to improve discriminative linguistic modelling working for Turkish broadcast newscast transcription. They used and compared 3 algorithms, specifically, MIRA, perceptron and SVM, both for ranking and classification. They applied the threshold parameter as a size minimization method on the scarce feature set, and several sample selection schemes to reduce the training complication. In this paper, they also extended the former outcomes by comprising the SVM classifier and classification and ranking versions of the MIRA algorithm for completeness. They also increased the feature space by incorporating higher-order -grams, presented the connection among ranking versions of perceptron and SVM, and gave a detailed statistical exploration and matching of the outcomes.

Pan and Li [100] have focused on FPGA based vigorous speech estimation and identification system, and the noise from the environmental issue was its key consideration. To speed up the identification and recognition speed of the FPGA-based SR system, the discrete hidden Markov model was employed here to reduce the computational complexity innate in SR. Moreover, the empirical mode decomposition was utilized to break down the quantified speech signal corrupted by noise into some intrinsic mode functions (IMFs). The IMFs were then weighted and added to recreate the original noiseless speech signal. Distinct from the earlier study, in which IMFs were carefully chosen by experimental and error for applications, the weights for each IMF were planned by the genetic algorithm to get an optimum way out. The investigational outcomes in this research study revealed that this technique achieved an improved SR rate for speech subject to many environmental noises.

Goh et al. [101] have proposed a speech recognition system utilizing harmonic structure-related information to sense harmonic features in a noisy environment. The proposed algorithm initially extracts the harmonic components enclosed inside the speech signals utilizing sine function convolution. By setting the frequency of the sine function as equal to the fundamental frequency of speech signals, harmonic components can be extracted. The reconstructed signal obtained by summing up the extracted harmonic components was found to have a high degree of correlation with the original signal. The extracted frame energy measure of the harmonic components has been further processed to become

dynamic harmonic features and then used together with the European Telecommunications Standards Institute (ETSI) front-end processed mel-frequency cepstral coefficients (MFCC) feature or the perceptual linear prediction (PLP) feature in the speech recognition system. The proposed enhanced speech recognition system showed a better recognition rate over the ETSI front-end processed MFCC (or PLP)-based speech recognition system.

Le et al. [102] have extended a new neural network language model (NNLM) which depends on word clustering to structure the output words: Structured Output Layer (SOUL) NNLM. This model was able to handle arbitrarily sized vocabularies, hence dispensing with the need for shortlists that are commonly used in NNLMs. Several softmax layers replaced the standard output layer in this model. The output structure depends on the word clustering which was based on the continuous word representation determined by the NNLM. Mandarin and Arabic data are used to evaluate the SOUL NNLM accuracy via speech-to-text experiments. Well-tuned speech-to-text systems serve as the baselines. The SOUL model achieved consistent improvements over a classical shortlist NNLM both in terms of perplexity and recognition accuracy for these two languages that are quite different in terms of their internal structure and recognition vocabulary size. An enhanced training scheme was proposed that allowed more data to be used at each training iteration of the neural network.

Ai et al. [103] have discussed the comparison of speech parameterization methods: Mel-frequency Cepstrum Coefficients (MFCC) and Linear Prediction Cepstrum Coefficients (LPCC) for recognizing stuttered events. Speech samples from UCLASS are used for analysis. The stuttered events are identified through manual segmentation and used for feature extraction. Two simple classifiers are used for testing the proposed features. The conventional validation method was used for testing the reliability of the classifier. The experimental investigation elucidated MFCC and LPCC features which can be used for identifying stuttered events and LPCC features were slightly outperformed MFCC features.

Rao et al. [104] have proposed global and local prosodic features extracted from sentences, words, and syllables for speech emotion or affect recognition. In this work, duration, pitch, and energy values are used to represent the prosodic information, for recognizing the emotions from speech. Global prosodic features represent the gross statistics such as mean, minimum, maximum, standard deviation, and slope of the prosodic contours. In this work, global and local prosodic features are analyzed separately and in combination at different

levels for the recognition of emotions. In this research paper, all the studies are carried out using simulated Telugu emotion speech corpus (IITKGP-SESC). These results are compared with the results of the internationally known Berlin emotion speech corpus (Emo-DB). Support vector machines are used to develop emotion recognition models. The results indicate that the recognition performance using local prosodic features was better compared to the performance of global prosodic features.

Ting et al. [71] have proposed a novel Self-Adjustable Neural Network in this paper with an application in speech recognition. SANN was able to adjust its network size according to the speech length and allows speech tokens with variable lengths to be trained and tested under the same network. SANN has been successfully used to recognize Malay vowels and 10 TIMIT isolated words. SANN outperformed the state-of-the-art HMM in recognizing Malay vowels, but its accuracy was lower in recognizing the TIMIT isolated words.

The existing techniques of other researchers are presented in this section. A model for Severe Speech Impairment has been developed by authors in [105]. The training model is composed of a small vocabulary and the speaker-dependent recognizer model with a reduced features extraction process. The results showed that when perplexity increases, the recognition rate increases. The analysis of redundant words is not focused on. In [106], the authors presented a Bayesian separation model for sparsity promotion and the perceptual wavelet domain. The wavelet coefficient is estimated from the discriminative information in the time-frequency plane. The result of this model was a degraded system with low SNR. The sensitivity of the speech degrades the performance of the hybrid classifier. In [106], the authors discussed the silent speech recognition model for Persons with Laryngectomy. The training of the features was done by using phrases with better running speech. Detecting when the speaker speaks and when he stops, is a major concern of error. Whisper is also a common issue that may exchange speech signals. Vector Taylor Series (VTS) is used for distinguishing the background from the whispering sound. Then, the model utterance defines its pseudo whisper samples. Cepstral features were developed frame by frame to derive whisper statistics. The analysis of transcribed whisper created complexity over desired targets.

The authors in [107] have presented an end-to-end deep learning model for speech reverberation and the acoustic modelling in which time-based DNN speech reverberation architecture to find the speech quality. Then, multi-channel microphone arrays were

developed for training the DNN based multi-conditioning. The application of non-linear approaches still has not resolved the classical DE reverberation problem. Further, biometric recognition using HTM spatial pooler and temporal memory. An array of circuit designs and examined over two datasets, namely, face recognition and speech recognition. Autocorrelation of each speech frame incurs design costs. Sampling rates and the delay in chip cause poor performance. In [108], the authors studied a couple of dictionaries for exemplar radar-based speech enhancement. Initially, full-resolution frequency between speech and noises is computed and stored as corpus dictionaries. Spectrogram features of a dictionary containing the exemplar help to achieve the full-resolution frequency domain. Based on the temporal context, the computational complexity of variant noise cases. Data feeding on lower-dimensional Mel features increases the mapping models.

In [109], the authors suggest a hybrid framework for developing a learning model using Hidden Markov Models. The log-likelihood score of each speech is computed and fed into a discriminative classifier. Finally, the deep neural network based HMM enhanced recognition accuracy. The intelligibility level of each uttered word operated on the choice of reference set degraded the accuracy. Imprecise and missing consonants found in these words of dysarthria speech lead to overlap among the respective classes. In [110], the authors studied Multi-Views Multi-Learners Approach using multi-nets ANN. The variability of the dysarthria speech was analyzed by the likelihood of vocabulary words and then fed into the learner's phases, both dependent and independent paradigms. Analysis of feature vector on associated words degrades the scalability and reliability due to incorrect classifications.

In [111], the authors present a piece of articulatory information for ASR systems. The study on the speech inversion process is analyzed in the form of vocal tract constriction variables. Aurora 2 speech corpus is utilized for training models. Then, word recognition tasks are performed using articulatory conjunction and enhanced the word recognition rates. Features extraction over noises must be focused on developing a robust environment. EMG based speech recognition using tackling speaking mode varieties like discrepancies between audible and silent speech have been developed. A spectral mapping algorithm is introduced between muscle contraction and word utterance. The presence or absence of phonetics makes a drastic difference. Maximum Entropy Gain (MEG) determines the

different signalling modes between audible and silent speech. Though it yields average computational complexity, silent speech analysis must improve.

Significance of speech manifolds analyzed over cross-lingual and multilingual by [112]. Manifold technique is studied for data selection and graph construction using deep neural networks. The representation of the features exhibited language interdependency. The results stated better usability whereas the analysis of inter-cross languages degrades the local structure of data. In [114], the authors studied Whispered Speech Recognition using deep noise encoders and estimated the measures like cepstral distances, cepstral coefficients, confusion matrix and inverse filtering which helps to classify the signals. The misclassification occurs due to train/test scenarios is of challenging tasks. In [115], the authors present speech enhancement by improvising full-sentence correlation and clean speech recognition. The effectivity of the speech segment and the speech utterances are estimated with predicted and unpredicted noises. The constrained maximization problem was resolved for different noises and overlapping data. Yet, the systems reduced data overlapping for small corpus.

In [116], the authors present an energy-efficient speech extraction process in mobile head-mounted display systems. The Fast Speech Extraction (FastSE) algorithm is adopted for the speech selection process which degraded the low latency speech extraction. A reconfigurable matrix operation accelerator is deployed for the energy-efficient process. Configurability and debugging of each network interface communication explored higher delay time. In [117], the authors presented a long span of temporal patterns in hierarchical deep network models. Here, the Link quality between posterior estimates of the ASR performance was studied. In [118], the authors studied acoustic coprocessors using the Hidden Markov Model (HMM) integrated 8 way -data path of NOR flash array and the Senones acoustic library. This system significantly reduced the error rate of 15.4% but the maintenance of an acoustic library is difficult. In [119], the analysis of regularized speaker adaptation using Kullback Leibler divergence based HMM estimated the posterior probabilities of each deep neural network-based acoustic model. The system performance degraded due to the instability of developing lexical models. In [100], the authors discussed embedded based recognition systems using empirical mode decomposition and the genetic algorithm. It reduced the computational load of the acoustic speech from Intrinsic Mode

Functions and further its estimated weight by using a genetic algorithm to obtain the optimal solution. It does not apply to real-time systems.

The authors in [120] deal with unknown unknowns from multi-stream speech recognition. Feedback based techniques are utilized for deriving out-of-vocabulary. By doing so, the behaviour of each signal is analyzed, and then relevant information is obtained. Then, the Gaussian mixture model is used as a classifier and thus speech signal was recognized. This system is mostly dominated by probability error. The study was enhanced by [121] using a biologically inspired learning system. Global communication in the neural network estimated the synaptic weight of its corresponding presynaptic and postsynaptic neurons. It reduced the overhead of the hardware implementation with the least recognition rate. In [122], the authors studied a joint approach model for single-channel speaker identification systems. This model supported the fully blind system followed by a minimum mean square error of sinusoidal parameters. Code vectors formulation from mixture estimation during the reconstruction stage degraded the signal quality.

In [123], the authors discuss the multilingual vision and the speech interaction process. The system estimated the eye gaze measure, head pose estimation, facial expression, and text-to-speech components and then fed into binary patterns of three orthogonal planes of the shape domain. It was experimented with in Art Critic and displayed better results. The time is taken by the number of feature points is higher. In [124], the authors presented a gating neural networks model for a large vocabulary audio-visual process. The assessment of visual features affects the performance of the system and it's resolved by introducing the gating layer [125] in neural networks. The system supports only 25% of the features for systems training. The integration of audio and visual features remains a challenging task during performance analysis. The fusion models are developed for the feature fusion model, the decision fusion model, the multistream hidden Markov model [125] (HMM), the coupled HMM, and the turbo decoders. Lattice generation at each time frame creates a constrained maximization problem. In [126], the authors studied recurrent neural networks for multi-genre broadcast speech models by linear hidden network adaptation layer and the k-component adaptive layers. Feature representation models helped to achieve better speech classification with degraded speech perplexity and word error rate.

2.6 CHAPTER SUMMARY

This chapter presents a view of existing automatic speech recognition methods. This literature review covered a variety of topics, techniques, methods, and approaches. The literature is categorized into four categories: the history of voice recognition systems, work carried out by other researchers in the field of speech recognition systems, performance analysis of speech recognition systems, and work related to proposed methodologies. The review presented in this chapter indicates the previously proposed methods of ASR for man-machine interaction. The usage of speech recognition systems is rapidly increasing in various engineering domains. The demand for robust speech recognition systems is being explored to a great extent in this thesis.

CHAPTER 3.

ISOLATED WORD RECOGNITION USING ANN

3.1 INTRODUCTION

The process through which a machine or a computer recognises spoken words is known as Speech Recognition. Simply, it can be said, it's a process in which the computer identifies what you are talking. SR techniques can be referred to as Computer speech recognition (CSR) or ASR. A computer program is required to implement an algorithm that generates word sequences through speech signals. Over the past sixty years, technological curiosity in the field of ASR attracted many researchers to automate the human dependent tasks into mechanical dependent tasks using humanoid and machine interaction.

Most of the study in speech processing is driven by the human desire to make mechanical models, so they can imitate the verbal communication capabilities of a human. The area of speech recognition has the fundamental objective to make means and methods aimed at speech input for machines [127]. Due to significant advancement in statistical speech modelling, the ASR methods have broad task application that requires human interface with the machine, it includes automatic processing of telephonic calls over networks and inquiry-based info methods to provide updated data about travel, and quotations of stock prices, reports on climatic changes etc. The signals in free space can be easily influenced by environmental noise like speech signals.

In the Automatic Speech Recognition method different type of environmental alteration remains a challenge, when the environment is noisy and compensating recognition is difficult. [128]. The speech signal is trained using some essential features and based on trained features the signal of speech is tested or known. Successively, the two major processes are testing and training the basic speech signal recognition (SSR) method. When the system performs in a noisy condition the throughput of the system degrades, which requires training and testing phase to improve the performance. To reduce this disparity and improve performance, numerous strategies have been created.

The feature enhancement method and model adaptation method are the two fundamental categories for grouping these methods. Remunerate the probability distributions of the recognizer directly using Model adaptation techniques and the extraction procedures are supportive in recognition of speech in the presence of noise. For significant speech

recognition short-time energy, pitch, formants, Mel-frequency range coefficients (MFCC), Teager energy operation and cross-section areas-based features are very fascinating.

At the test time, the received denoising feature vectors can function as a feature recognition approach and these denoising features are trained from very clear speech such that they go with the recognizer acoustic model [129]. As compared to model domain techniques these methods are more attractive due to simpler computation and the recognizer can implement them independently. For example, some methods such as spectral subtraction work at the spectrum level, and spectral-MMSE works on the features specifically. Usually, to assist the process of enhancement a priori speech model is employed by the Gaussian mixture model (GMM). For instance, SPLICE works on stereo recordings of noise-free and noise interrupted speech to understand the piecewise linear mapping of interrupted speech to clean speech using a Gaussian mixture model.

While front-end methods have shown improved performance on some tasks make point estimates of the clean speech features. Errors in estimates that further result in degradation of performance is due to the mismatch between acoustic models and features. Better execution and recognising of accurateness in raucous conditions can be enhanced by these methods.

3.2 PROPOSED AUTOMATIC SPEECH RECOGNITION SYSTEM

The simple aim of an ASR system is to improve functioning in human-machine interaction. It inspires the suggested Automatic Speech Recognition scheme to identify the received message/signal and translate the determined audio message/signal to the equivalent message/signal.

The proposed system comprises four stages which include pre-processing, extraction of features, selection of optimum features then identification. The structural design of the suggested Automatic Speech Recognition scheme is presented in figure 3.1.

Pre-processing the feed-in recorded acoustic signal for noise elimination and to identify the word. The next step is to extract the features, namely, pitch, word size, and sampling point. These are the physical features. And five analytical features like Entropy, Variance, Skewness, Mean and Kurtosis are also extracted. Consequently, the optimum degree of features for the classifier's training session is chosen.

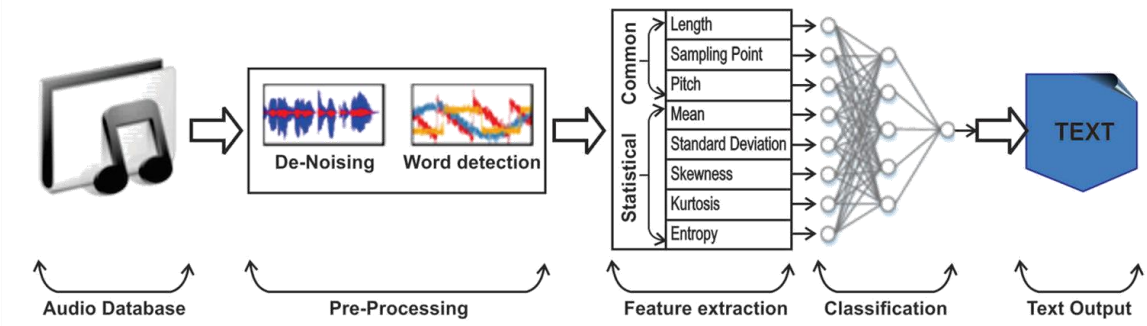


Figure 3.1: Structural design of Proposed ASR system

The next step is to identify the pronounced word and shows the text matching it. The following sections present a detailed explanation of the proposed ASR system.

3.2.1 Input Data collection

The input data collection comprises speech signals spoken and collected from different people. In the presented ASR system, this collection of speech signals is treated as the input/received signal and mathematical expression 3.1 denotes the received speech signal written below,

$$S_i = x_i(t), i = 1,2 \dots N \tag{3.1}$$

In the above equation 'S_i' represents the i^{th} input signal from the collected data and 'N' signifies the complete number of speech signals in the collected data.

3.2.2 Pre-processing

In the ASR method, the first phase is pre-processing which manages time-varying analog speech signals at the time of recording. To further process the signal in the digital domain, sampling of the continuous-time signal is done to convert the speech signal into a digital signal (discrete-time discrete-valued). Partition the speech into a series of unrelated segments or frames as the characteristics of a signal vary gradually in time and each segment is treated the way it has permanent characteristics. In this supposition, anyone can take out the features of each segment derived from the tester within the frame. On further processing, the real signal will be replaced by a feature vector which means an analogue speech signal that changes with time, is converted into a series of feature vectors. The procedure of changing series or segments of samples of speech to feature vectors signifying events in the probability space is termed Signal Modelling.

Pre-processing has the function to obtain a group of factors introduced as of the transmission channel in an arrangement to denote speech signals beneficial for successive dealing out and to execute the sampled version of the speech signal and create illustration free of variations in amplitude, speaker noise and stress. The pre-processing step uses both the time and frequency domain approaches, among this Time dominion methodologies are commonly simple to apply and directly deal with the speech signal waveform with factors of energy and zero-crossing rates. Frequency dominion methodologies include certain kinds of spectral investigation not directly observable in the time dominion. The Frequency dominion methodologies are greatly utilized in SR systems.

Background Noise Removal:

In actual conditions, background noise is generally generated from ACs, fluorescent lamps, fans, PCs, typewriters, the sound of foot stepping, talkback, traffic, noises from doors etc., the designers of SR techniques often have litter regulations over these things. Additive noise is typically fixed in nature excluding impulse noise causes such as typewriters. Based on the surroundings, ranging from 60dB to 90dB the noise degrees will vary. A head-mounted close speaking microphone is the usually adapted method to minimalize the impact of the background noise. When an utterer generates speech at a normal conversation level and the microphone passes speech signal through a filter process at that time the average speech level rise by approximately 3dB on every occasion. Equation 3.2 gives the filter employed to eliminate the background noise.

$$E_s = 10 * log_{10}\left\{\in + \frac{1}{N}\sum_{n-1}^{N} S^2(n)\right\} \quad (3.2)$$

In the above equation 'E_s' represents the log energy of a block of 'N' samples and '$\in$' is a small positive constant added to avoid the calculating of log zero. 'S (n)' represents the n^{th} speech sample in the block of N samples.

Speech Word Detection:

The speech recognition needs to execute the sound comprising of speech, quietness/silence and various noises. The identification of the existence of speech inserted in the dissimilar sounds and noises is recognized as detection of speech or endpoint detection or speech activity recognition. Because of several advantages, a good endpoint detection algorithm impacts the execution of the system for speech and accuracy. Firstly, remove the silent segment before recognition and for each, speech and noise, the amassed sound likelihood

count focuses extra on the utterance part of the speech. Secondly, it becomes hard to represent noise in addition to silence accurately in changing environments and limit this effect using the background noise segment beforehand. Thirdly, eliminating non-speech segments, if existing in large numbers can considerably minimize the computational period.

For the detection of speech words following steps are required:

Step 1*: C*omputation for the detection of the endpoint: Zero crossing score, NZ represents the score of zero-crossing the block. In illustration 3.2 Es symbolises the log energy of a block of length N samples. The 3.3 equation symbolises the regularization auto-correlation coefficient at unit sample delay C_1.

$$C_1 = \frac{\sum_{n=1}^{N} S(n).S(n-1)}{\sqrt{[\sum_{n=1}^{N} S^2(n)]*[\sum_{n=0}^{N-1} S^2(n)]}} \tag{3.3}$$

Step 2: Filtering for detection of the endpoint: Suppose one sound might have probable silences which isolate many speech segments. Identify a set of endpoints called beginning segment and endpoints to find every segment. There is dependably a leading-edge ensuing a starting point and a falling edge afore final point on the energy forms of utterance. At first, the approach is to sense the edge following this detect the corresponding finish points or endpoints. Require a detector to identify every feasible endpoint from energy features for robust and precise detection of an endpoint. Equation 3.4 specifies the detection of the endpoint has the characteristic for 1-D provisional energy in the sample data as specified below.

$$E(l) = 10 \times log10 \sum_{j=n(l)}^{n(l)+l-1} o(j) \tag{3.4}$$

In the above equation, '$o(j)$' is the data sample, 'l' is window length, '$E(l)$' is frame energy in decibel, '$n(l)$' is the number of first data sample in the window.

Step 3: Energy normalization: The reason behind energy normalization is to regularize the E_l i.e., the utterance energy. Equation 3.5 specifies the energy normalization for obtaining $E_{\max}$ i.e., the maximum energy value over the words as,

$$E_{\max} = \max(E_l), \quad 1 \leq l \leq L \tag{3.5}$$

Subtracting $E_{\max}$ from E_l to give $\hat{E}_l = E_l - E$. In this manner, the highest energy value of every word is 0 dB, and the identification system is comparatively unaffected by the

variance in gain among different recordings. During the abovementioned computations, there exists a constraint that word energy from regularization cannot happen before locating the end of the word.

3.2.3 Extraction of features

Extraction of features is a crucial process in the ASR method, providing the exact recognition of words depending on the system features. This approach enhances the sharpness of the identified speech depending on the selection of features and feature extraction. Herein suggested Automatic Speech Recognition system, the work is exceptionally focused on feature extraction. The common features like pitch, word length, sampling point and statistical features like Entropy, Variance, Skewness, Mean and Kurtosis are considered.

(i) Mean: It is the arithmetician's terminology for the average estimate of a signal symbolised as μ and generated in general, by adding together all the samples, and dividing by N. Equation 3.6 represents these as a mathematical form.

$$\mu = \frac{1}{N} \sum_{i=0}^{N-1} x_i \tag{3.6}$$

(ii) Standard deviation: It is alike the average deviation, except the averaging, occurs with power instead of amplitude. Accomplish these by taking the average on squaring each deviation (remember, power $\propto$ voltage^2). To complete, take the square root to compensate for the beginning square. Equation 3.7 represents an equation form.

$$\sigma = \frac{1}{N-1} \sum_{i=0}^{N-1} (x_i - \mu)^2 \tag{3.7}$$

(iii) Skewness: It is defined as an estimate of the unsymmetrical distribution of the probability of random variables with real values approximately it is mean in statistics and probability theory. It can hold either negative or positive value, or even indefinite value. In equation 3.8 the moment's estimator for N sample values, a regular approach for population skewness is defined.

$$\gamma = \frac{\frac{1}{N-1} \sum_{i=0}^{N-1} (x_i - \mu)^3}{\sigma^3} \tag{3.8}$$

(iv) Kurtosis: It is defined as an estimate of the "peakedness" of the distribution of the probability of a random variable with real value in statistics and probability theory. With the same idea of skewness, kurtosis is a figure/pattern with the form of a distribution of

probability and distinct approaches for assessing it for a theoretic distribution and correlating approaches for approximating it from a sample of the population, just like skewness. Several explanations of kurtosis are mainly tail weight, peakedness (width of peak), and absence of shoulders deduce the exact estimates. Using equation 3.9 calculate the Kurtosis as

$$Kurt = \frac{\frac{1}{N-1}\sum_{i=0}^{N-1}(x_i-\mu)^4}{\sigma^4} \tag{3.9}$$

(v) The speech signal entropy: It is essential to measure the differential entropy of a process or a method with certain surveillance in different commerce arenas consisting of unconstrained component analysis, image analysis, genomic study, SR, manifold learning, and delay in time approximation. In this research work, the modest and very generally used method, a histogram-based estimation is applied. Equation 3.10 gives the expression used to compute the histogram-based entropy estimation.

$$Entropy = -\sum_{i=0}^{N-1} x_i log\left(\frac{x_i}{w(x_1)}\right) \tag{3.10}$$

In the above equation, w represents the width of the i^{th} bin.

3.2.4 Identification of Word

At this stage, convert the recognized uttered words into text. The accuracy of the identification highly relies on the classification method. Penetrate artificial intelligence (AI) to tackle this problem. In this research work, the AI algorithm used for classification is an artificial neural network (ANN).

3.2.4.1 ANN Based Word Recognition

The principal aim of the programmed calculation framework, ANN is to match the neural network and working of the humanoid brain. It is composed of an intertwined network of theatrically formed neurons that work as a route for the exchange of data. ANNs are versatile and adjustable, learning with all dissimilar external or internal stimuli. ANNs are utilized in pattern recognition systems and sequence detection, processing of data, modelling and automation. The Artificial Neural Networks include a single input and output layer and single or multiple hidden layers. Excluding the input layer, compose each node with neurons. Each layer consists of a different number of nodes based on the issue.

The number of hidden layers and nodes defines the complexity of the architecture. Training an Artificial Neural Network is to determine a group of weights that provide preferred values at the output when fed with different patterns at its input. Training and testing are the key practice of an ANN. Figure 3.2 demonstrates a sample of a simple ANN.

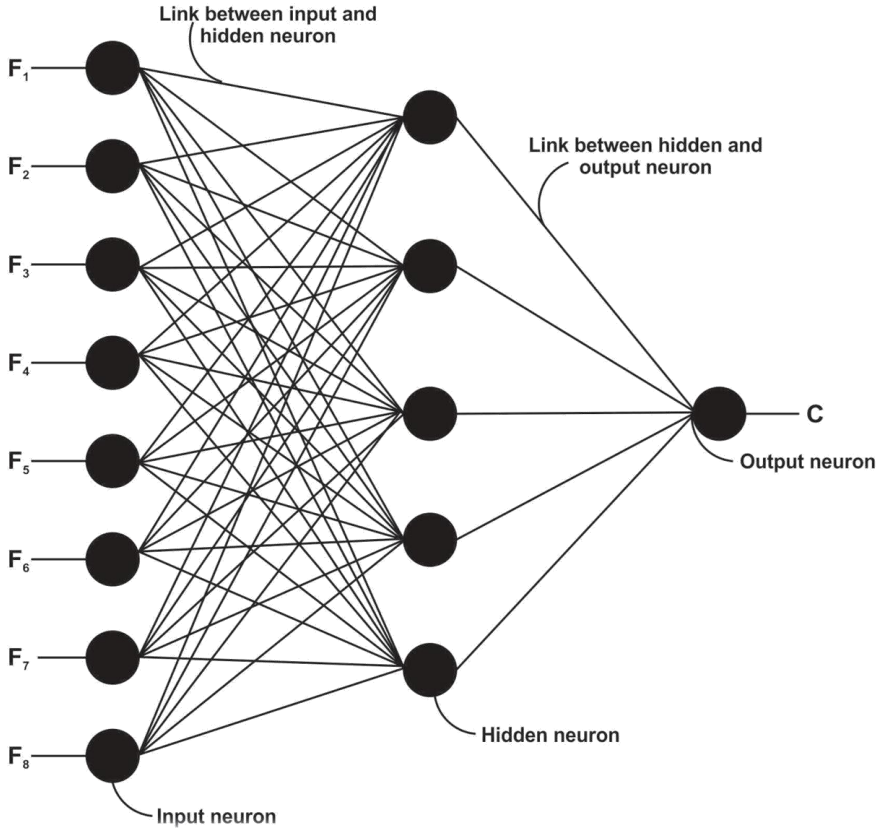


Figure 3.2: Artificial Neural Network Structure

F_1 to F_8 represents eight features that are input to the ANN and the corresponding text can be obtained as the output of the ANN. The eight features include five Statistical features and three common features. Mean, skewness, standard deviation, and kurtosis are the statistical features and pitch, word length and sampling point are the common features. Hence, the suggested ANN structural design comprises eight inputs and conforming word output. Training and testing are the two key methods of classification algorithms.

At this stage, the extracted features are exploited to train the Artificial Neural Network. In training, define the input, and output and then set the suitable weight to facilitate the ANN classifier to guess the appropriate word in the testing phase. In any classifier algorithm, the training phase plays the main role. For training, the suggested system uses the

backpropagation (BP) algorithm. The procedure incorporated in the training is specified below and figure 3.3 illustrates the structural design of the BPNN.

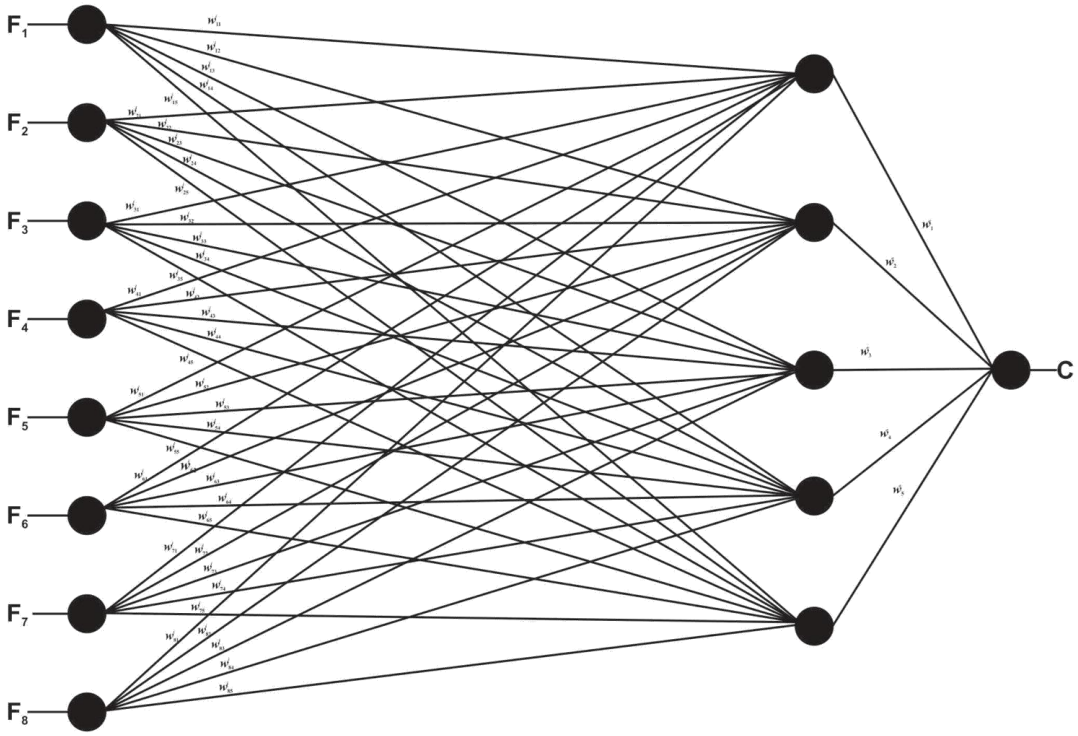


Figure 3.3: Suggested BPNN

The suggested Artificial Neural Network consists of 8 input components, 1 output component, and 5 (M) hidden elements. In the forward pass of the BP algorithm, First, transmit the input data to the hidden layer and then, to the output layer. Every node in the hidden layer gains input through the input layer and is interconnected with appropriate weights and added. Achieve the output of the NN using equation 3.11 stated underneath.

$$C = \sum_{j=1}^{M} \frac{w_j^O}{1+\exp(-\sum_{i=1}^{N} F_i w_{ij}^I)} \tag{3.11}$$

In equation (3.11), 'F_i' represents the i^{th} input value and the weights allotted amid hidden and output layer is given as 'w_j^O', the weight allotted amid input and the hidden layer is given as 'w_{ij}^I' and the number of hidden neurons is given as M. The outcome from the hidden node is the non-linear transformation of the resultant summation. The output layer follows the same process. Compares the output values from the output layer with target

values and calculates the learning error rate for the NN, which is specified in equation 3.12 as:

$$\partial_k = \frac{1}{2}(Y - C)^2 \tag{3.12}$$

Here, ∂_k represents k^{th} the learning error of the Artificial Neural Network, Y is the preferred output and C is the real output. The backward pass of the BP algorithm transmits the error among the nodes headed back to the hidden layer. Then, repeat the training for the new training database by altering the weights of the NN.

3.2.4.2 Reduction of error by the BP algorithm.

(i) First, assign the weights to the neurons of the hidden layer. The input layer holds a fixed weight, and randomly chooses the weights for the neurons of the output layer. Next, calculate the output via the mathematical expression given in equation 3.11.

(ii) Then, compute the back-propagation error using equation 3.13.

$$BP_{error} = \sum_{k=1}^{Z} \partial_k \tag{3.13}$$

In the above equation 'BP_{error}' represents Back Propagation error, learning error rate of k^{th} training data set is represented as '∂_k'. Utilizing equation 3.14 found the weight deviation in the hidden neuron.

$$\Delta w = BP_{error} \,.\, \gamma \,.\, \delta \tag{3.14}$$

In the above equation weight deviation is symbolised as 'Δw', 'γ' symbolises learning rate, which generally varies from 0.2 to 0.5, 'δ' symbolises the average of hidden neurons output.

$$\delta = \frac{1}{\mathrm{T}} \sum_{n=1}^{T} H_n \tag{3.15}$$

Here,

$$H_n = \frac{1}{1+\exp\left(-\sum_{i=1}^{N} F_i w_{in}^{I}\right)} \tag{3.16}$$

Where 'δ' represents the average output of hidden neurons. The overall number of input neurons is given as 'N', overall training is given as 'T' and activation function at the input side is given as 'H_h' and it is h^{th} output at the hidden neuron. Next, utilizing equation 3.17 the new weights are found.

$$w^{new} = w + \Delta w \tag{3.17}$$

In the above equation 'w^{new}' represents the updated weight or new weight and 'w' represents the current weight.

Then redo the procedure till the Back Propagation error becomes reduced and it assures $BP_{error} < 0.1$. ANN becomes prepared for classification when the BP error attains the least value.

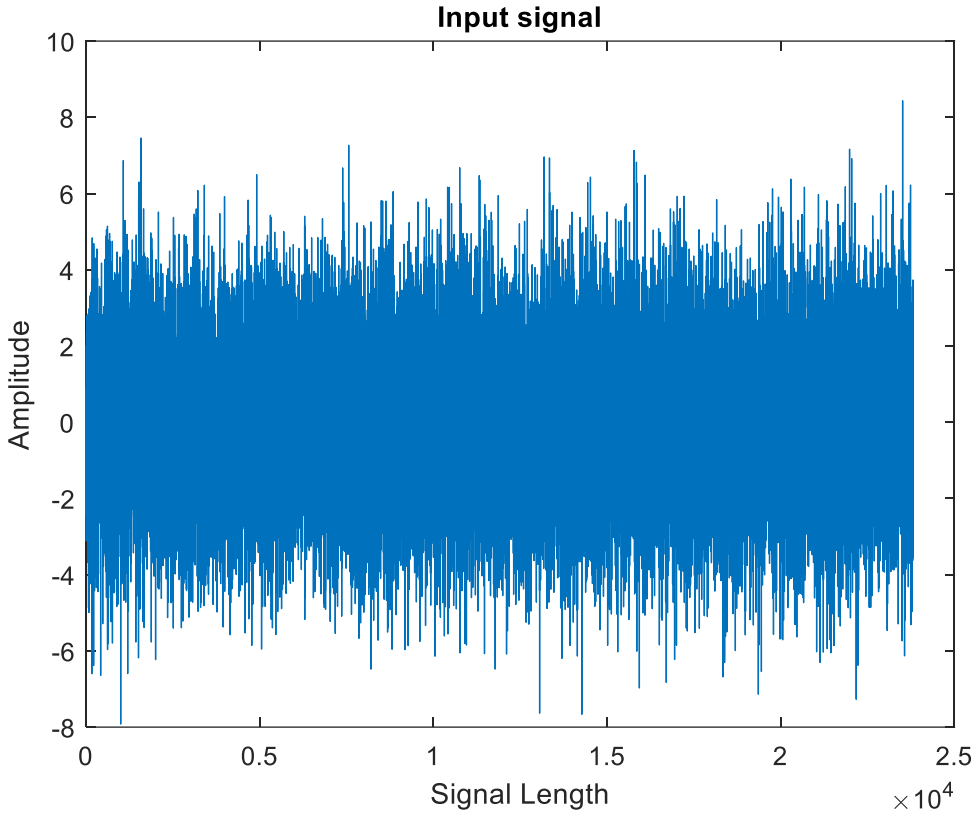
Next, process the testing, and the classification accuracy of ANN is verified using the testing process. During this process, the classifier validates a different and new set of data except for the data castoff in the training. As soon as the ANN fulfils the criteria of accuracy for classification it can be used for the practical implementation of the planned classification.

3.3 EXPERIMENTAL RESULT AND DISCUSSION

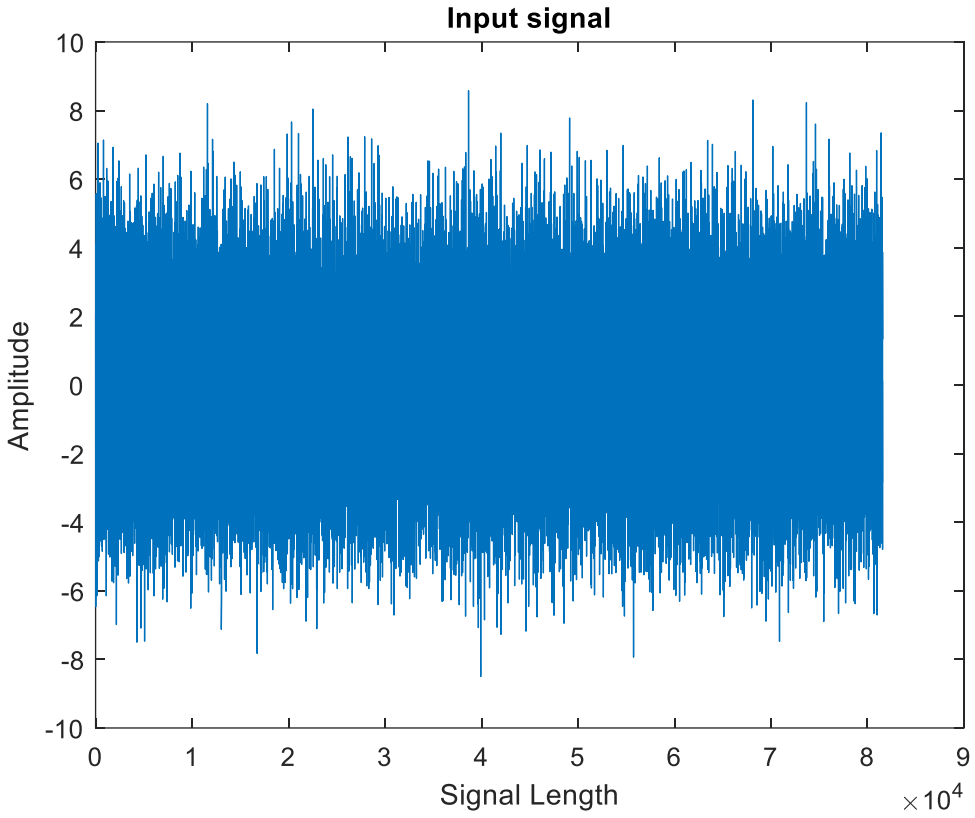
The suggested system for ASR is applied in the functioning platform of MATLAB with the given system description.

Processor	:	Intel i5@3GHz
RAM	:	8GB
Operating System	:	Windows 7
MATLAB Version	:	R2013a

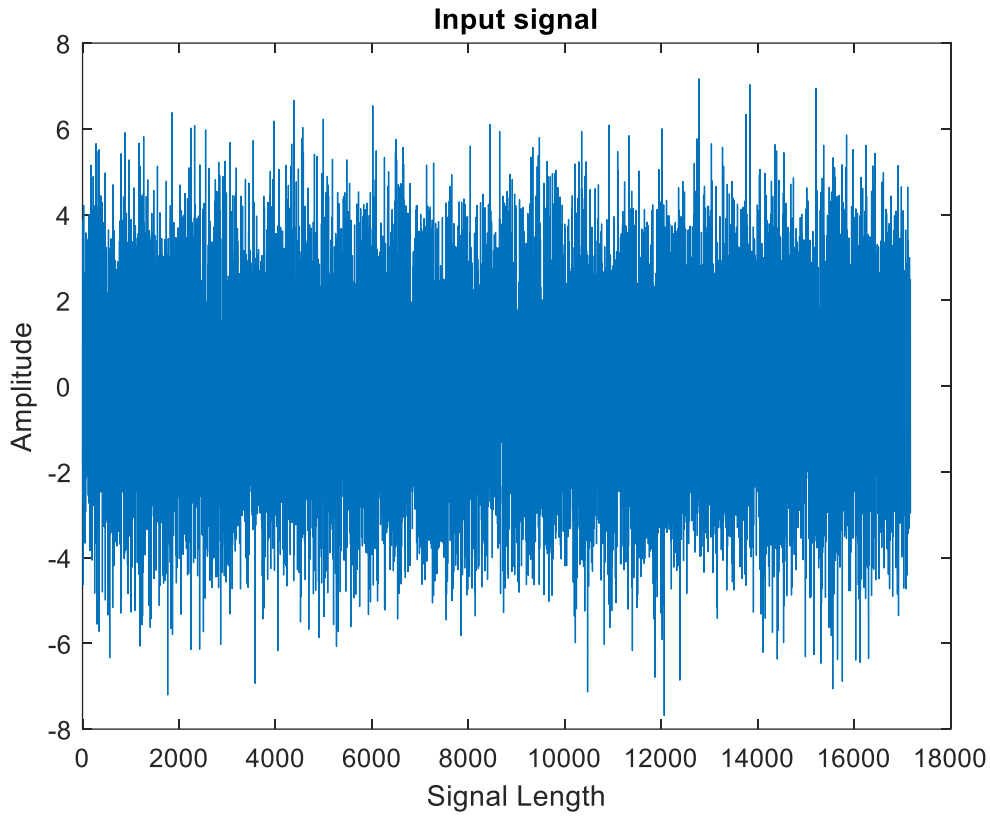
A specific method to create interaction between humans and machine is the ASR system. With the help of artificial intelligence, machine and human interaction become possible. Anyone can control or interact with the machine, but speech is very special and the simplest way the interaction among various attitudes. In this case, recognizing the spoken word or speech accurately a speech recognition system is essential. Subsequently, the machine can understand the speech of an ordinary man by transforming these identified speeches or words into an equivalent electrical signal. In this research work, an ASR system to identify 5the pronounced word and transform it into the equivalent text is developed. In this proposed system, MATLAB is the research tool for implementing the system.



(a) Input speech signal 'Apple'



(b)Input speech signal 'Badam'



(c)Input speech signal 'Cow'

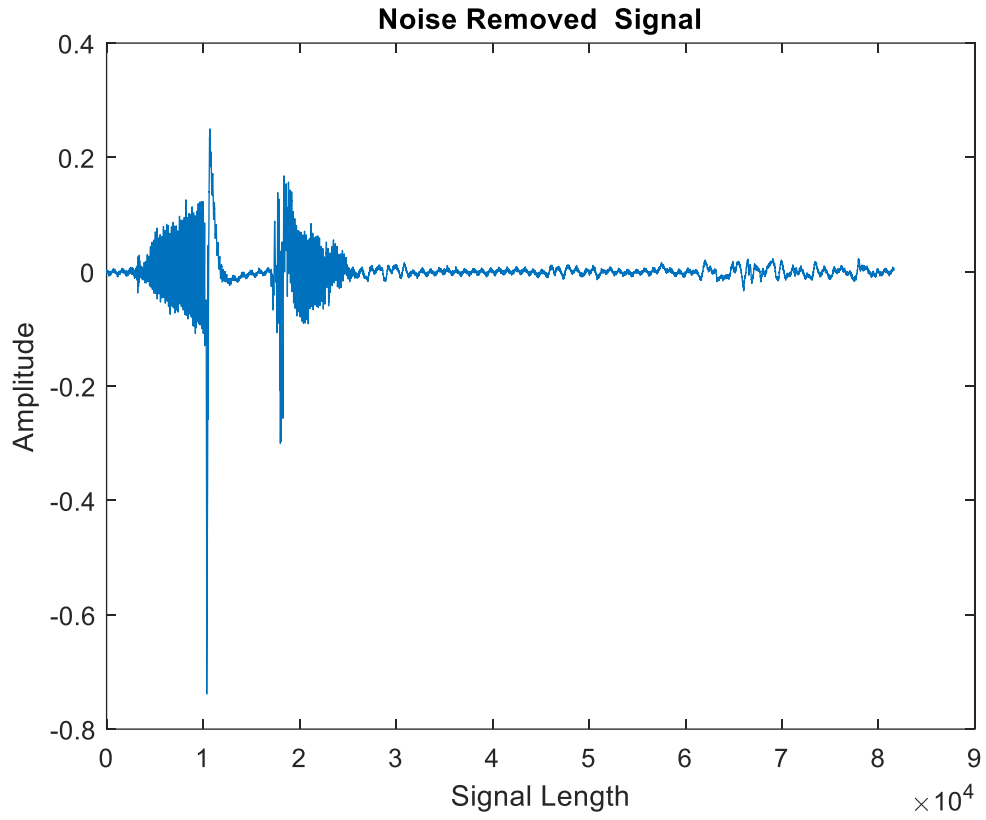
Figure 3.4: Input Speech Signals (a) Apple, (b) Badam (c) Cow

Analyse the recognition performance using the recorded speech signals. The database consists of three hundred forty-six recorded speech signals. They are collected and stored in the input audio database. Among 346, three signals (Apple, Cow and Badam) have been used for the performance analysis. Figure 3.4 shows the input speech signals: Apple, Badam and Cow.

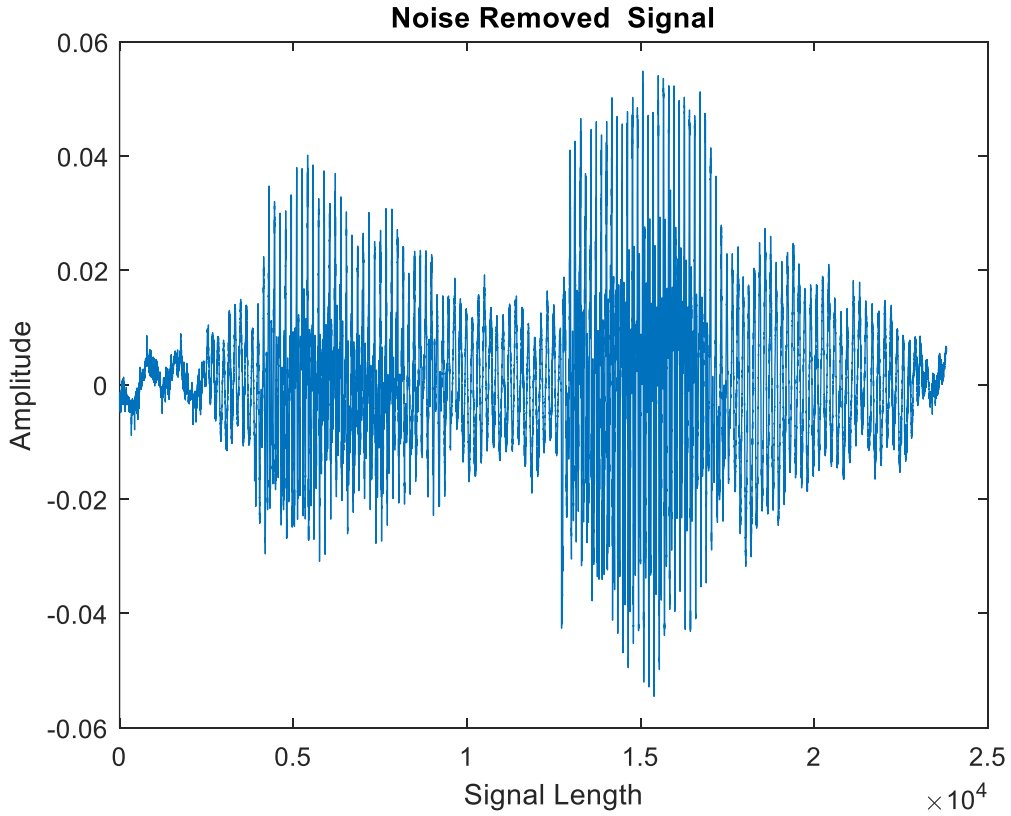
This input speech signal is preprocessed to remove noise. The noise removed signals for the three speech signals, which are selected for performance analysis are shown in figure 3.5. Also, the detected word signal for three words that are used for the analysis is shown in figure 3.6. The three words used from a database of 346 words are:

(a) APPLE
(b) BADAM
(c) COW

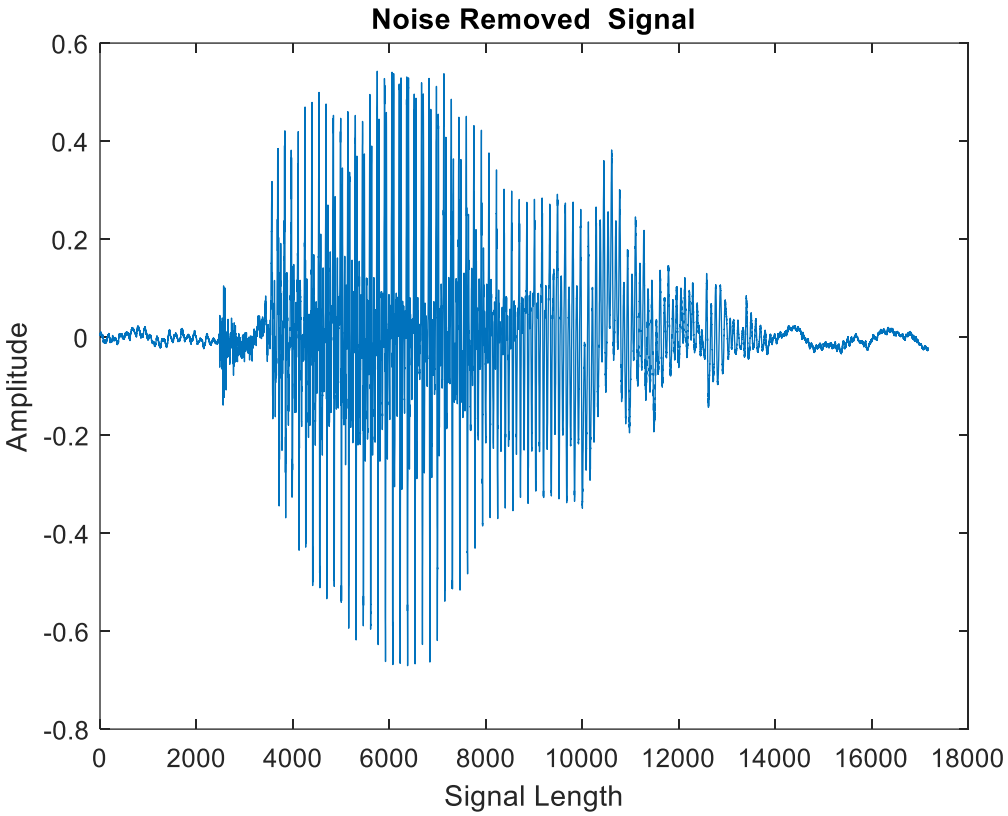
The Preprocessed denoised signal and detected word signal are shown in the following graphs for these three words. These are amplitude vs signal length graphs.



(a): Pre-processed De-noised signal 'Apple'

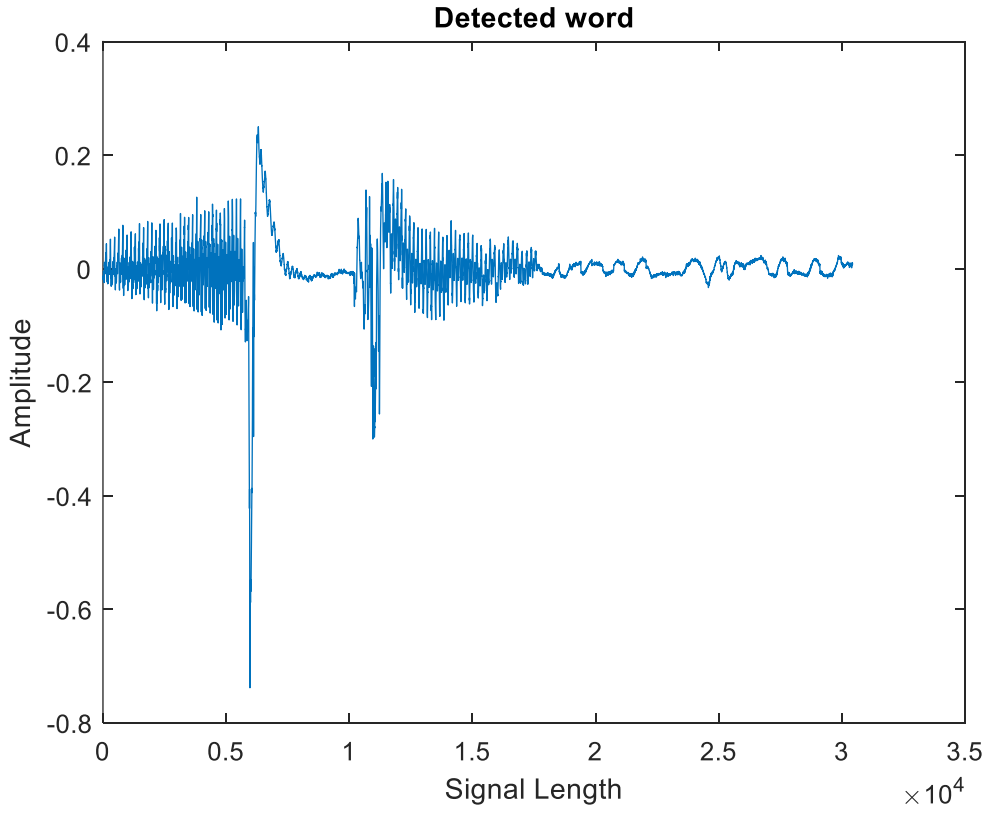


(b): Pre-processed De-noised signal 'Badam'

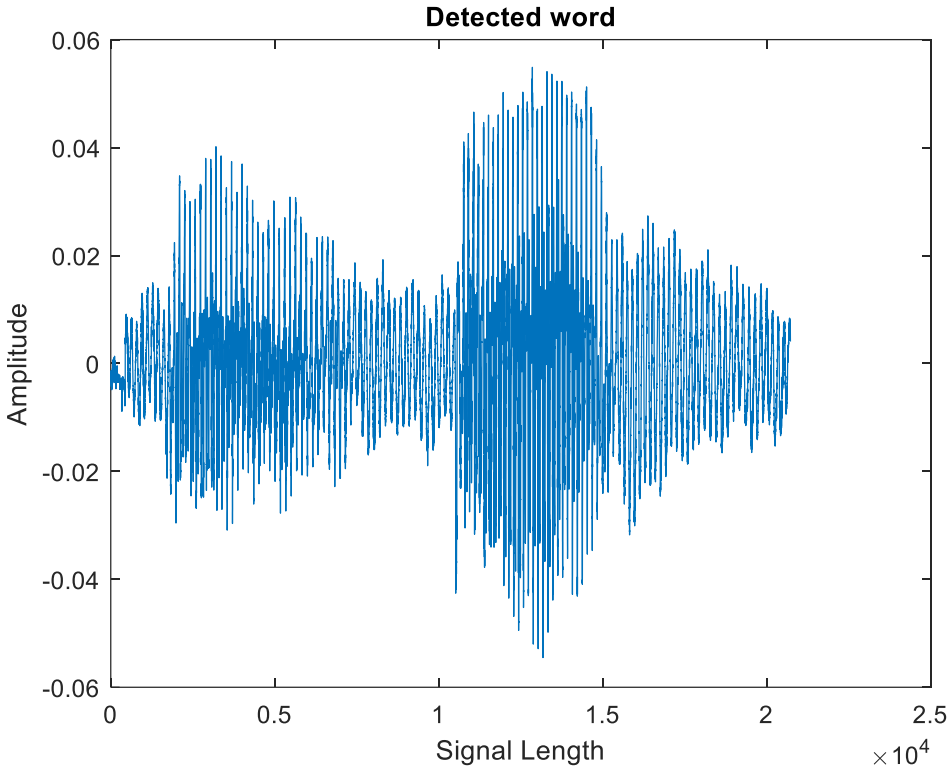


(c): Pre-processed De-noised signal 'Cow'

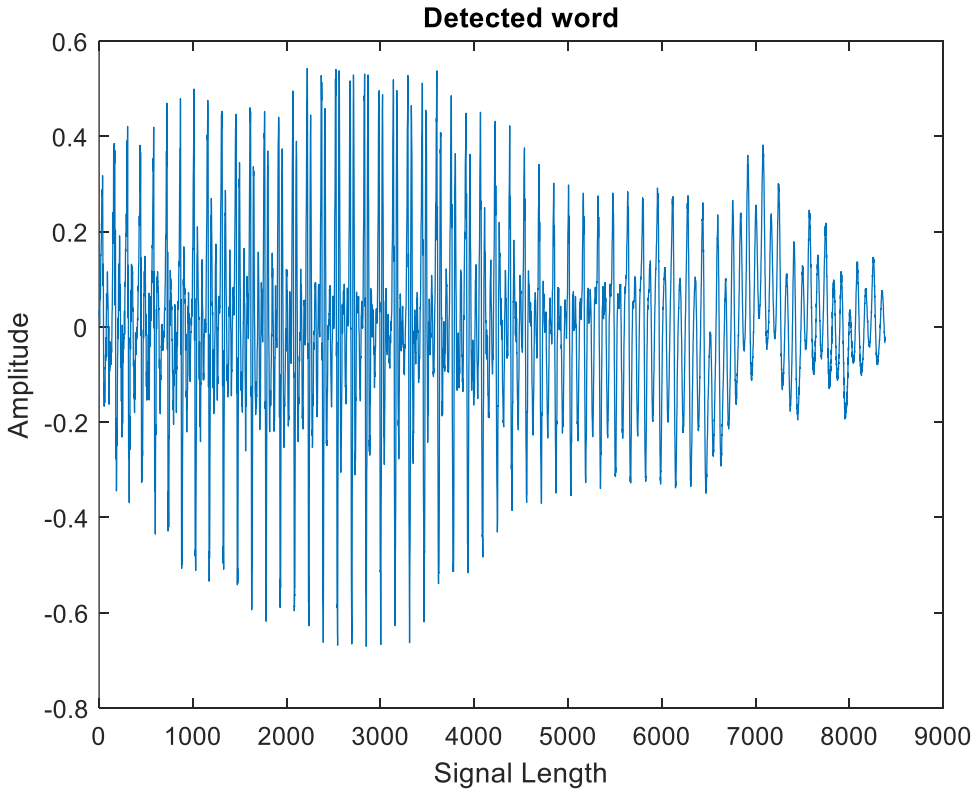
Figure 3.5: Pre-processed De-noised signal (a) Apple, (b) Badam, (c) Cow



(a): Pre-processed Detected word signal 'Apple'



(b): Pre-processed Detected word signal 'Badam'



(c): Pre-processed Detected word signal 'Cow'

Figure 3.6: Pre-processed Detected word signal (a) Apple, (b) Badam, (c)Cow

Figure 3.5 shows the pre-processed signal for 3 diverse speech signals: Apple, Cow and Badam. The pre-processing step occurs for eliminating noise and identifying the isolated word. After this pre-processing stage, the statistical features and the other general features of the uttered signal are extracted. The common features are easily obtained after pre-processing. The statistical features were obtained using equations 3.6 to 3.10.

The statistical features so obtained are enlisted in Table 3.1.

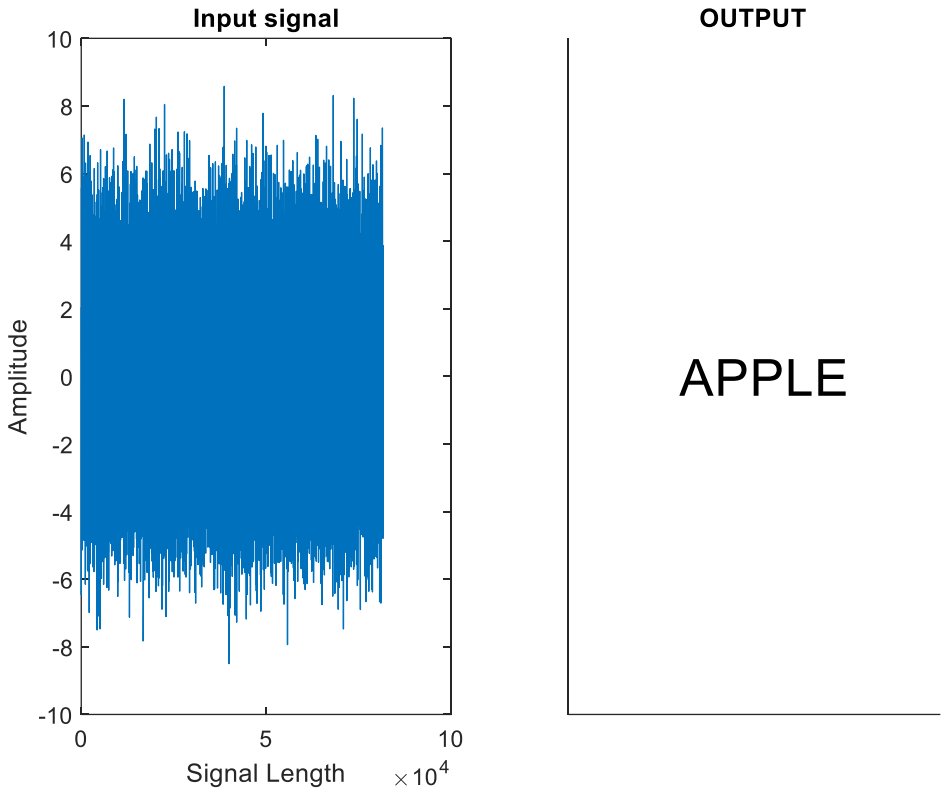
Table 3.1: Attained Statistical features values

Features	***Numerical value of Signal Samples***		
	Apple	***Badam***	***Cow***
Mean	-0.0000077	0.0022	-0.00033
Variance	0.0011	0.0549	0.0173
Skewness	-0.00000023	-0.0000087	-0.0000057
Kurtosis	0.00000000017	0.000028	0.00000071
Entropy	5683.3	15510	5242.4

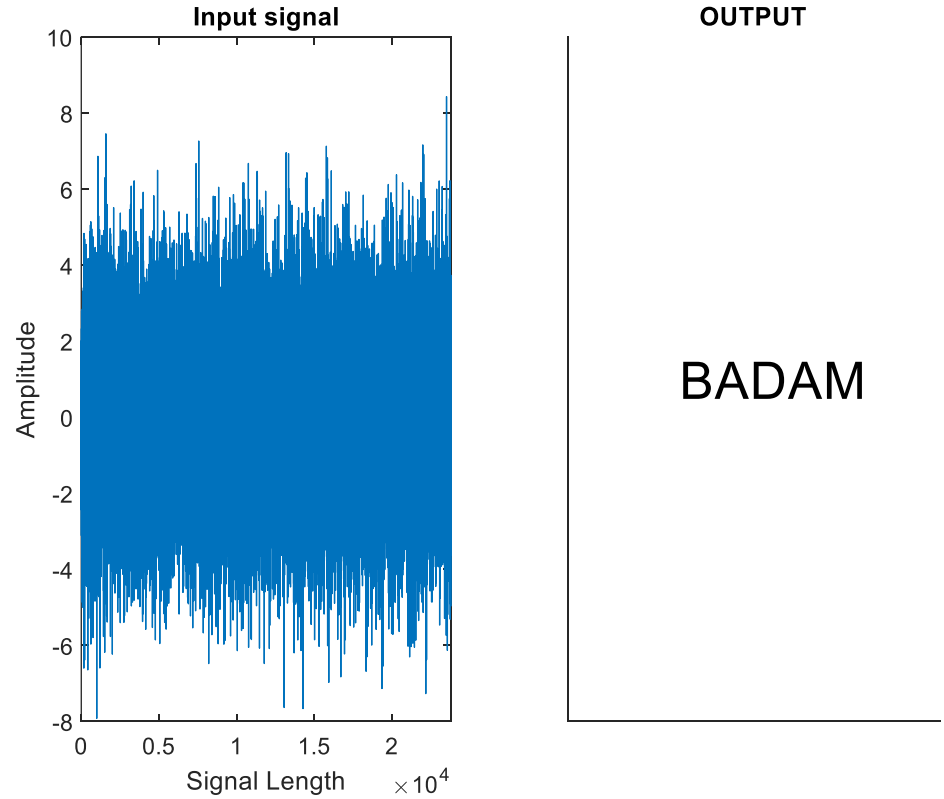
The 8 features together with statistical and common features are fed to the Artificial Neural Network for the classification of the signal. The classification after matching leads to displaying the corresponding text.

Figure 3.6 illustrates the pre-processed detected word signal for input speech signals.

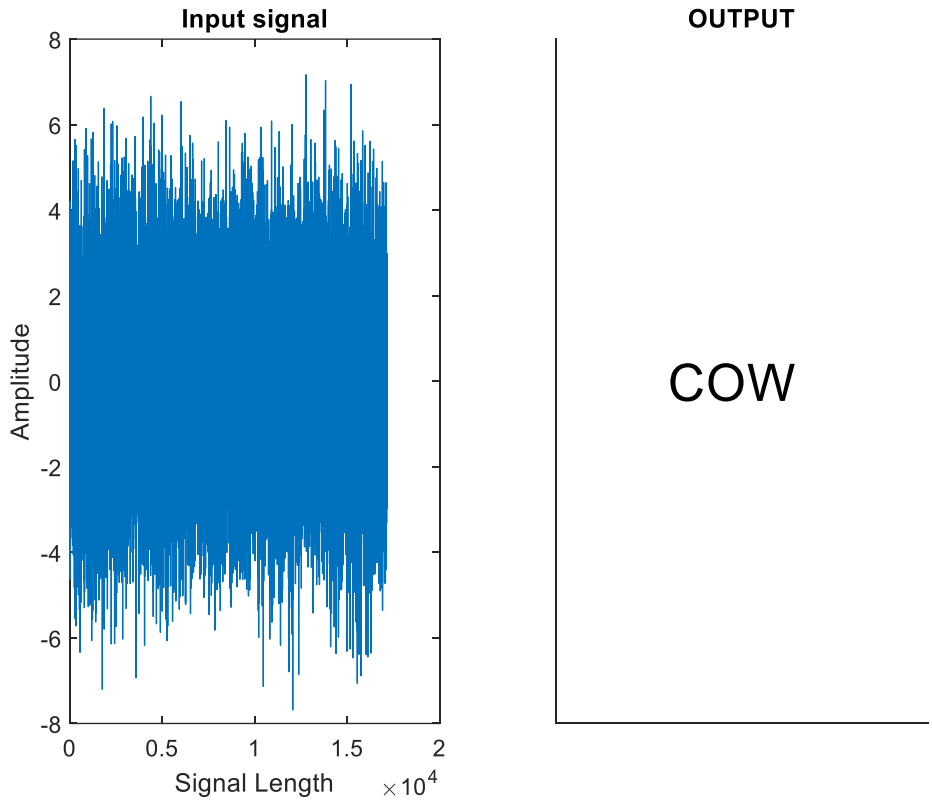
Figure 3.7 shows the attained outcome of the suggested ASR system. The outcome of the ASR system is the matching text of the uttered word. It is displayed on the screen as shown in figures 3.7 (a), (b) and (c).



(a): Output Text display 'Apple'



(b): Output Text display 'Badam'



(c) Output Text display, 'Cow'

Figure 3.7: Output Text display (a) Apple, (b) Badam, (c) Cow

The performance analysis of the Artificial Neural Network classifier is based on the classification Accuracy, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), False Positive Rate (FPR), False Discovery Rate (FDR) and Matthews Correlation Coefficient (MCC).

Figure 3.7 shows the graphical representation of these performance measures. The figure illustrates the performance of the classifier of the suggested automatic speech recognition system. The performance obtained for Sensitivity, Specificity, Accuracy, FPR, PPV, NPV, FDR and MCC are 50%, 74%, 62%, 26%, 65%, 60%, 35% and 0.19% correspondingly. These obtained outputs show that the suggested ASR system can fulfil the necessary factors such as sensitivity, accuracy and specificity for recognizing the uttered word.

From these results and performance analyses, it is demonstrated that the suggested or presented system is a fit and the appropriate method for Automatic Speech Recognition tasks.

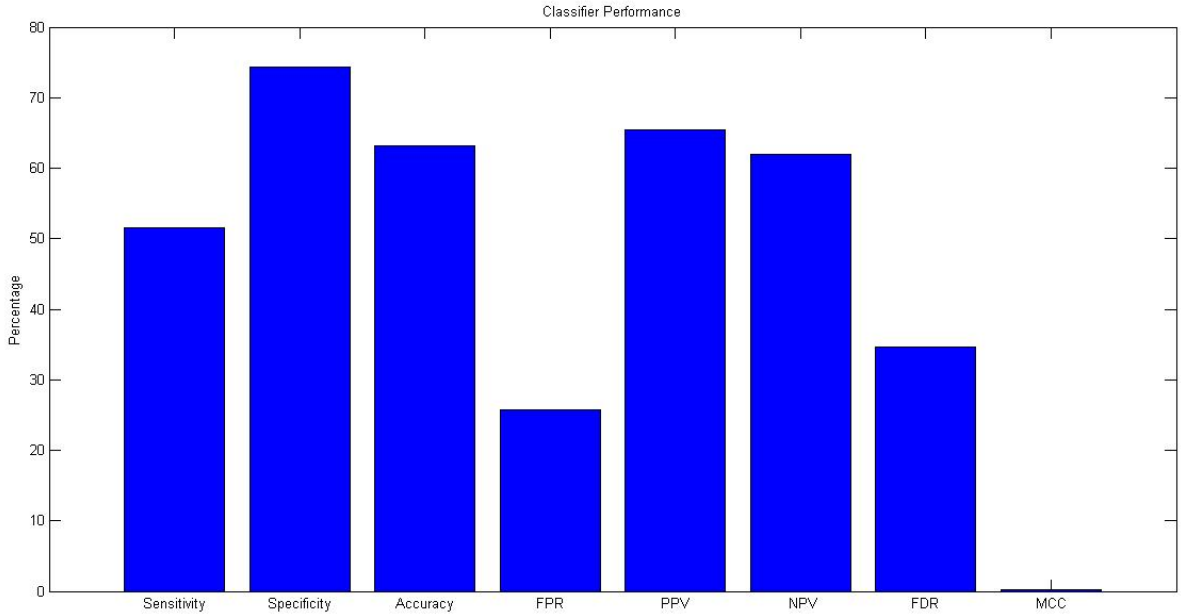


Figure 3.7: Classification Performance of proposed ASR

3.4 CHAPTER SUMMARY

The proposed ASR system converts the spoken speech signal into the corresponding text. This ASR system consists of three stages; at the initial pre-processing stage of the uttered speech signal to remove noise and to detect the isolated word is performed. In the second stage of feature extraction, the eight features including word size, sampling frequency, Skewness, Mean, Variance, pitch, Entropy, and Kurtosis are drawn from the signal. In the last phase, a spoken word gets identified with the support of extracted features and displays the respective text. The word identification occurs using an ANN classifier with back-propagation training. Then the working of the proposed system for 3 samples of signals like Apple, Cow and Badam is described and the conforming waveforms are displayed. The evaluation of the classifier relies on the accuracy of the classifier, sensitivity and specificity are determined. The overall performance shows that the suggested approach for Automatic Speech Recognition is the finest option for the human-machine interface. It is suggested that the presented ASR scheme provides improved performance, so it is apt for an actual man and machine interface through speech. In the future, the performance of the ASR system will be improved by the effect of a unique classification approach.

CHAPTER 4.

OPTIMAL SELECTION OF FEATURES BASED ON HYBRID ABC-PSO

4.1 INTRODUCTION

Speech Recognition is the ability of a machine or program to identify words and phrases from spoken language and convert them into a machine-readable format. It is also known as Automatic Speech Recognition or computer speech recognition and speech to text conversion. Speech recognition frameworks have turned into one of the chief applications for machine learning and pattern recognition innovation [130]. Speech recognition is an extremely troublesome undertaking to be performed by a computer framework. This circumstance is because of the variability in the way individuals talk which brings about complex speech signals that must be processed via automatic speech signals. They must be handled via automatic speech recognition systems (ASRS). In speech recognition, there are several key areas of examination for the present advancement of spoken language frameworks. These key regions are automatic speech recognition, robust speech recognition, unconstrained speech, and so on. Numerous consumer and modern applications oblige quick and lightweight real-time speech recognition with restricted vocabulary, for example, without hands control for compact music players, car sound frameworks, cordless telephones, and residential applications. Speech recognition is a pattern classification issue and speech recognition frameworks utilize detached word recognition.

Current speech recognition frameworks utilize a pattern matching methodology. The classifier, which is ordinarily considering hidden Markov models (HMM). Hidden Markov model (HMM)-based speech recognition advances have grown extensively and can now get a high recognition execution. Voice dictation frameworks, spoken dialogue frameworks, and speech data interfaces are illustrative speech applications that utilise these modern innovations. These advancements lead us to anticipate that speech input interfaces will be inserted in practical applications. The improvement of speech input interfaces implanted in versatile terminals obliges recognition precision, scaling down, and low-power consumption [132]. Past examination of custom equipment portrayed the execution of the HMM calculation utilizing application-specific integrated circuits (ASICs) and field-programmable gate arrays (FPGAs).

There are several sorts of parametric representations for acoustic signals. Among them, the Mel-Frequency Cepstrum Coefficients (MFCC) is the most broadly utilized. There are numerous reported works with MFCC, particularly on the improvement of the recognition accuracy.

The utilization of Mel Frequency Cepstral Coefficients can be considered one of the standard techniques for feature extraction. The utilization of around 20 MFCC coefficients is basic in ASR, although 10-12 coefficients are frequently thought to be adequate for coding speech [133]. The most prominent drawback of utilizing MFCC is its sensitivity to noise because of its dependence on the spectral structure. Strategies that use data in the periodicity of speech signals could be utilized to overcome this issue, even though speech likewise contains an aperiodic substance. It has been observed that feature extraction algorithms and classification methods perform an important role in the area of stuttered events recognition.

4.2 PROPOSED METHODOLOGY

The objective of this research is to present an efficient recognizer using speech recognition techniques to produce an interface between humans and machines. This proposed technique consists of a few steps including preprocessing, feature extraction, and optimal feature selection for recognition. The preprocessing is done to increase the effectiveness of the feature extraction process. Feature extraction is to extract the features of the given speech signal while selecting the most optimal features.

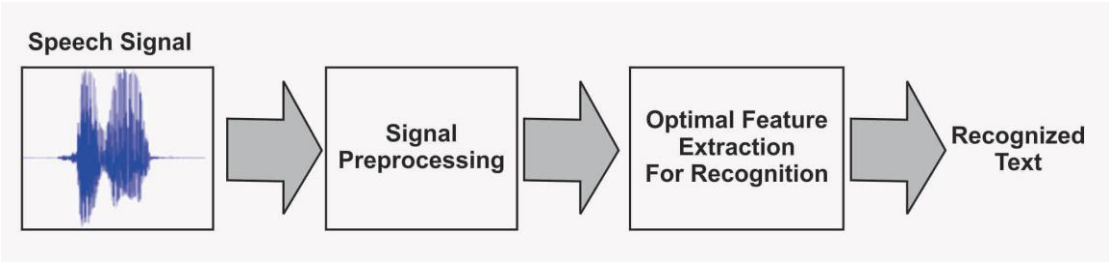


Figure 4.1: Process flow of Proposed System

To select the most optimal set of extracted features a feature selection algorithm is implemented. Then finally these optimal features for used for the recognition of text. The process flow of the proposed automatic speech recognition system is shown in figure 4.1.

4.2.1 Preprocessing

The input speech signals have noise in it which will influence the recognition process. Consequently, the first imperative step in speech recognition is the preprocessing of the speech signals or acoustic signals which is executed to remove avoidable waveform of signal and to shorten the task of recognition. In this preprocessing, the Wiener filter is applied to evacuate the noise which deals with the standard of spectral subtraction. This diminishes the noise by assessing the noiseless signal and afterwards comparing it with original signals. It takes up that noise is stationary in examination with the non-stationary signal in this way subtracting it from the original signal. After this process, signals are pre-emphasized to normalize the word counters by diminishing the high spectral dynamic range. The signal went through high pass FIR with a distinctive component value. This procedure is finished by detecting the endpoints of the signal and evacuating the silence. Finally, a noise removed is obtained at the preprocessing step without the loss of original data which is given to the feature extraction process.

4.2.2 Optimal Feature Extraction and Selection Based on Hybrid ABC-PSO

For speech recognition, from the pre-processed speech signal it is needed to extract the features for recognition of text. In this proposed method, eight types of features are extracted they are statistical features like mean, variance, standard deviation, skewness, energy, and acoustic features like pitch, MFCC, and LPCC. The statistical features of the pre-processed speech signal are mentioned below.

Consider that signal consists of n samples as y_1, y_2, y_3… y_n then,

(i) Mean: The mean of the sample is denoted by μ, and it is calculated by using the below formula which is given by

$$\mu = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{4.1}$$

(ii) Variance: Variance is the arithmetic mean of the squared deviations from the sample mean which is given by

$$\text{Variance, } V = \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{n-1} \tag{4.2}$$

(iii) Standard Deviation: The standard deviation is related to the average deviation; the only difference is that averaging arises with power instead of amplitude. It is given by

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{n-1}} \tag{4.3}$$

(iv) Skewness: Skewness is a measure of the degree of asymmetry of a distribution. The skewness of the sample signal is calculated by using

$$\gamma = \frac{\frac{1}{n}\sum_{i=0}^{n}(y_i-\mu)^3}{SD^3} \tag{4.4}$$

(v) Energy: The energy of the speech signal is given by

$$E = \sum_{i=0}^{n} Y^2(n) \tag{4.5}$$

Some of the acoustic features of the pre-processed speech signal are mentioned below.

(vi) Pitch: It signifies the apparent fundamental frequency of a sound. It is one of the major auditory qualities of sounds along with loudness.

(vii) Mel Frequency Cepstrum Coefficients (MFCC): It is well known and most widely used feature extraction technique for speech recognition. A Mel is expressed as a unit of measure of the ear's apparent frequency. The approximation of Mel frequency is expressed as

$$Mel(f) = 2595 \times \log\left(1 + {}^{f}/_{700}\right) \tag{4.6}$$

Where $Mel(f)$ indicates apparent frequency and f indicates the real frequency. Initially, the pre-processed speech signal is divided into frames and each frame is windowed with some window function to reduce the discontinuities of the speech signal by narrowing the start and end of each frame to 0. The window function should have low sidelobe levels and a narrow main lobe in the transfer function. Here, the window function is performed by using the hamming window. Then fast Fourier transform (FFT) block converts frames from time to frequency domain. Next, Mel scaled filter bank is used to convert real frequency to apparent frequency called the Mel frequency scale. The Mel frequency warping is usually realized by three-sided filter banks with the middle frequency of the filter. In the next stage, the log of filter bank output is calculated and finally, discrete cosine transform (DCT) is calculated. MFCC is calculated by using the below equation which is given by

$$C_s = \sum_{k=1}^{K}(\log S_k)\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right] \tag{4.7}$$

Where n= 1, 2... K.

K represents the number of Mel cepstrum coefficients.

(viii) Linear Predictive Coding Coefficients (LPCC): The LPCC analysis of a speech sample can be estimated as a linear grouping of previous speech samples. LPCC is a frame-

based analysis of the speech signal which is executed to provide observation vectors of the speech sample. To calculate LPCC features, first, the speech signal is divided into frames of samples. Every frame is multiplied by a sample Hamming window, and this windowed frame is delivered to perform short term autocorrelation. Then, LPCC analysis is implemented based on the Levinson-Durbin recursion. It gives a 2Qby-T matrix of observed features. The LPCC coefficients are then transformed into Q cepstral coefficients, which are biased by a raised sine window. The first half of an observed vector is the weighted cepstral sequence for a particular frame, the second part is the time difference weighted cepstral coefficients which are utilized to add dynamic information. Here, 24 feature vectors are extracted using LPC analysis.

From these eight extracted features it is needed to select the most optimal set of features for the recognition of speech signals. Artificial Bee Colony (ABC) and Particle Swarm Optimization (PSO) algorithms are well-known evolutionary algorithms that will mimic the insect or bird's behaviour in problem modelling and solution. Here, a combined Artificial Bee Colony and Particle Swarm Optimization algorithm is used to select the most optimal set of features extracted in the feature extraction process.

Steps involved in the ABC-PSO algorithm:

1. Initialize the extracted features from the speech signal as a food source position.
2. Each employed bee yields a new food source in their food source site and exploits the better source.
3. Each onlooker bee chooses a source depending on the quantity of her solution, produces a new food source in their food site and exploits the better source.
4. Determine the source to be abandoned and allocate its employed bee as a scout for searching for new food sources.

   After this step, the PSO algorithm is used in place of scout bees for searching for new food sources.
5. Initialize the population of new food sources as particles with random positions.
6. Calculate the fitness value for the given objective function for each particle. The fitness function for choosing the optimal set of features is based on equivalent reduced feature space and predefined classification accuracy. The fitness function is given by

$$Fitness = \alpha\psi_s + \beta.\frac{|N|-|S|}{|N|} \tag{4.8}$$

Where ψ_s classifier performance for the subset S, N is the number of total features, β is feature subset length and α is the classification quality.

7. Set present particles as "Pbest".
8. Add velocity to initial particles to obtain a new set of particles.
9. Find fitness value for each new set of particles.
10. Compare each particle's fitness value to find a new "Pbest" between the two sets of particles.
11. Find the minimum fitness value by comparing two sets of particles and the corresponding particle is "Gbest".
12. Update velocity for the next iteration using the below formula,

$$V^{k+1} = V^k + a(P_{best} - x^k) + b(G_{best} - x^k) \quad (4.9)$$

$$x^{k+1} = x^k + V^{k+1} \quad (4.10)$$

13. The iteration of PSO is repeated until the convergence is made.
14. Memorize the best food source found so far.
15. Repeat steps 2- 14 until the stopping criterion is met.

The classification performance for selecting the most optimal set of features for recognition is done by Support Vector Machine (SVM). SVM is established as a non-probabilistic binary linear classifier, where the optimal decision hyperplane is computed by support vectors. When the optimal feature vectors are not linearly distinguishable, a kernel function SVM is selected to map the vectors into a new feature space in which converted vectors can be classified by the optimal decision hyperplane. The training set comprises positive and negative training vectors which are considered as x. For the speech recognition technique, the speech feature vectors are represented by x=+1 and non-speech feature vectors as x=−1. The SVM decision function is given as:

$$f(y) = sgn(\sum_{i=1}^{L} \alpha_i \, x_i K(y_i, y) + b) \quad (4.11)$$

Where L denotes the number of support vectors, y_l is *l*-th support vector (feature vector), and α_i is weight allocated to the *l*-th support vector with its label given by x_i and b is a constant bias. K (.) is the kernel function that accomplishes implicit mapping from the unique feature space to a higher dimension feature space. After the SVM model is trained, the testing feature vector y will be classified according to the SVM decision function f(y), where it is allocated to the class +1 if f(y)≥1 and to the class −1 else. The architecture for the overall proposed system is shown in figure 4.2.

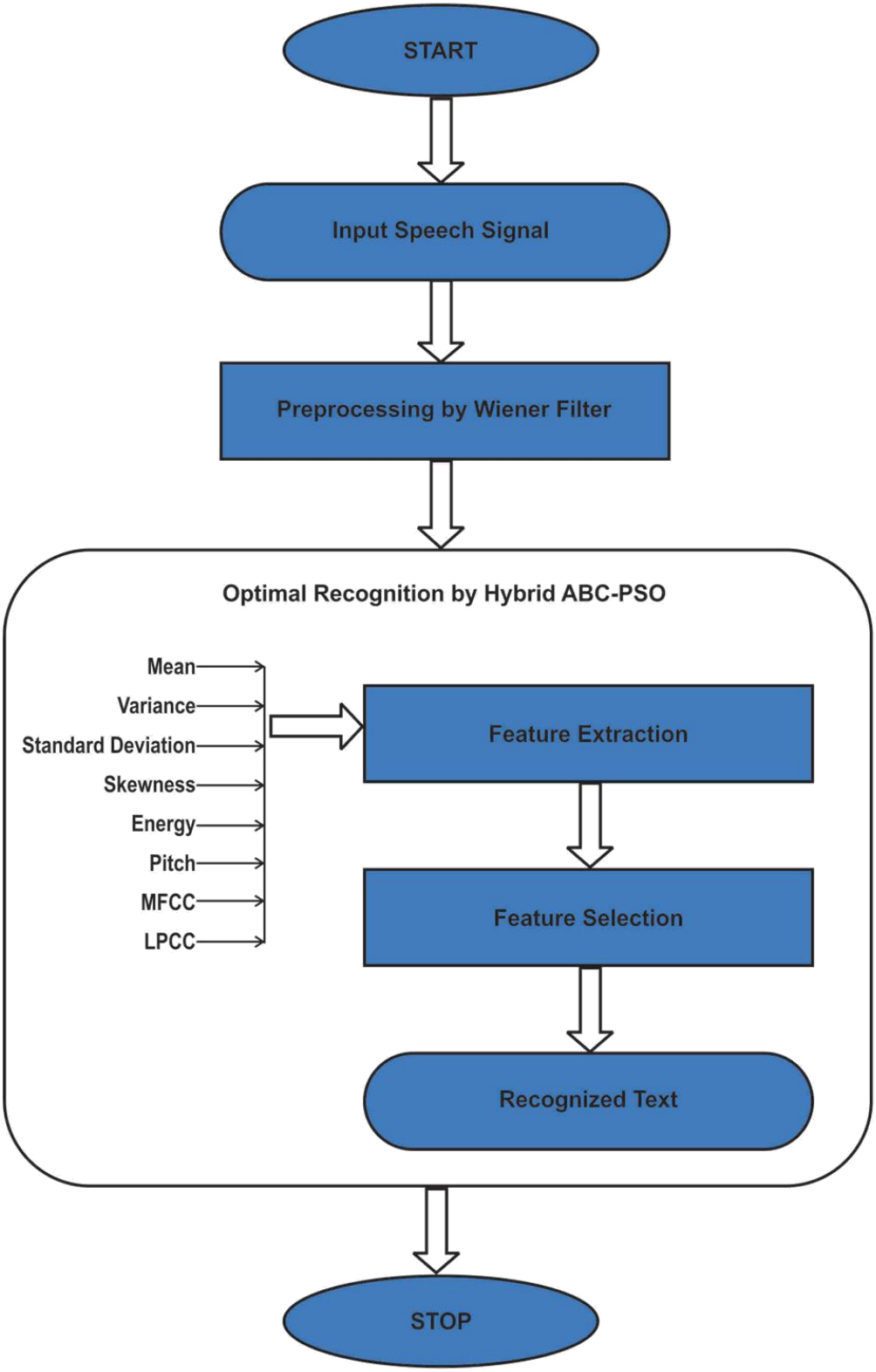


Figure 4.2: Architecture of Overall Proposed System

From, the architecture it can be seen that the proposed system recognizes the speech signal for recognition. Initially, the speech signal is pre-processed by the Wiener filter and the

pre-processed signal is given to the Hybrid ABC-PSO algorithm. In that phase, eight sets of features are extracted and then based on the classification accuracy optimal set of features is selected for speech recognition. Finally, the system produces recognized text for the given input signal.

4.3 EXPERIMENTAL RESULTS

The proposed system for automatic speech recognition is implemented in the working platform of MATLAB with the following system specification.

Processor : Intel i5 @ 3GHz
RAM : 8GB
Operating system : Windows 8
MATLAB version : R2013a

The speech data set comprises 100 different words in which these words were spoken continuously and recorded using high-quality microphones under noise conditions. The dataset was recorded by using Audacity 1.3 beta. The training data comprises 3045 utterances spoken 30 males and 26 females. All the speakers are natives of India, educated in the age group of 18 to 35 years. From the dataset, one speech signal is chosen and used for analysis. Figure 4.3 shows the input speech signal taken from the dataset which is given to the preprocessing stage.

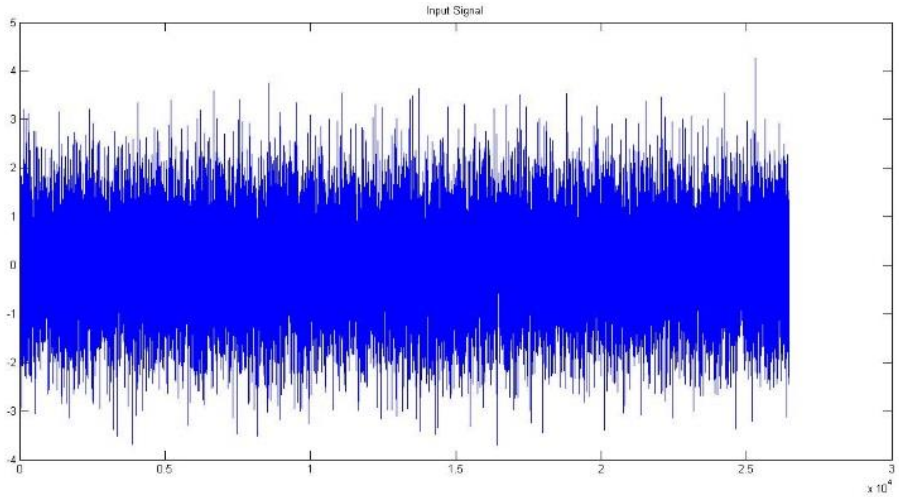


Figure 4.3: Noisy input speech signal

By applying the wiener filter to this noisy signal noise removed signal is obtained which is shown in figure 4.4.

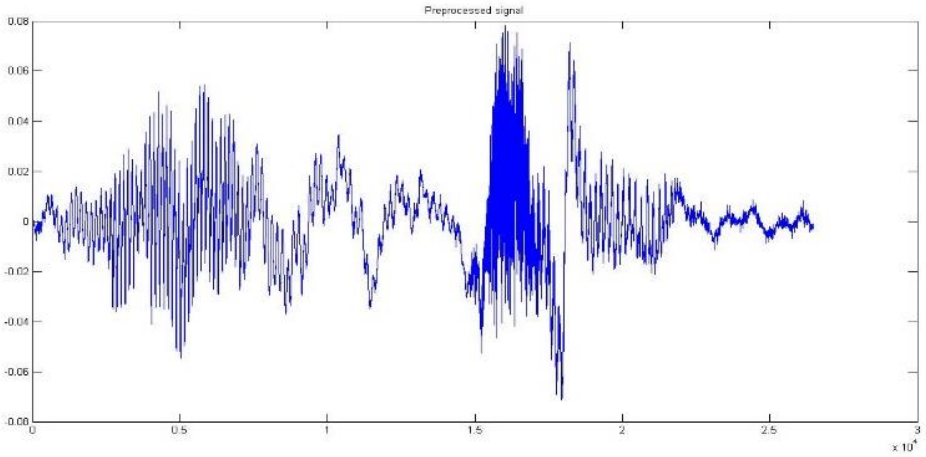


Figure 4.4: Pre-processed Speech Signal

In the feature extraction process, the statistical and acoustic features of the pre-processed signal are calculated, and their feature values are given in table 4.1.

Table 4.1: Feature Values

S. No.	Features	Values
1	Mean	0.01023
2	Variance	0.0111
3	Standard Deviation	0.1239
4	Skewness	-0.05378
5	Energy	7965.88
6	Pitch	198.02
7	MFCC	0.01778
8	LPCC	0.12987

Next hybrid ABC-PSO is implemented to select the most optimal features obtained from the feature extraction process. Here a total of 312 features are extracted from the speech signal. At first 100 numbers of maximum iterations applied for extracted 312 features are

applied. At that point, the optimized finest feature subset is computed. The above methodology is performed for 200 and 300 iterations moreover. Table 4.2 demonstrates the optimized finest feature subset for the greatest number of iterations 100, 200 and 300 for the extracted features.

Table 4.2: Optimized finest feature subset

No. of Iterations	Total Features	Optimized finest feature subset
100	312	256
200	312	213
300	312	187

From table 4.2, it is seen that when the number of iterations gets increased, the length of the features subset gets reduced. A Speech recognition system consists of a training stage and a testing stage. In the training stage, the SVM models are prepared for every speaker. In the testing stage, the stored information is compared with the demanded SVM model, and a choice is made. Figure 4.5 shows the obtained output of the proposed ASR system. The output of the ASR system is the corresponding text of the spoken word.

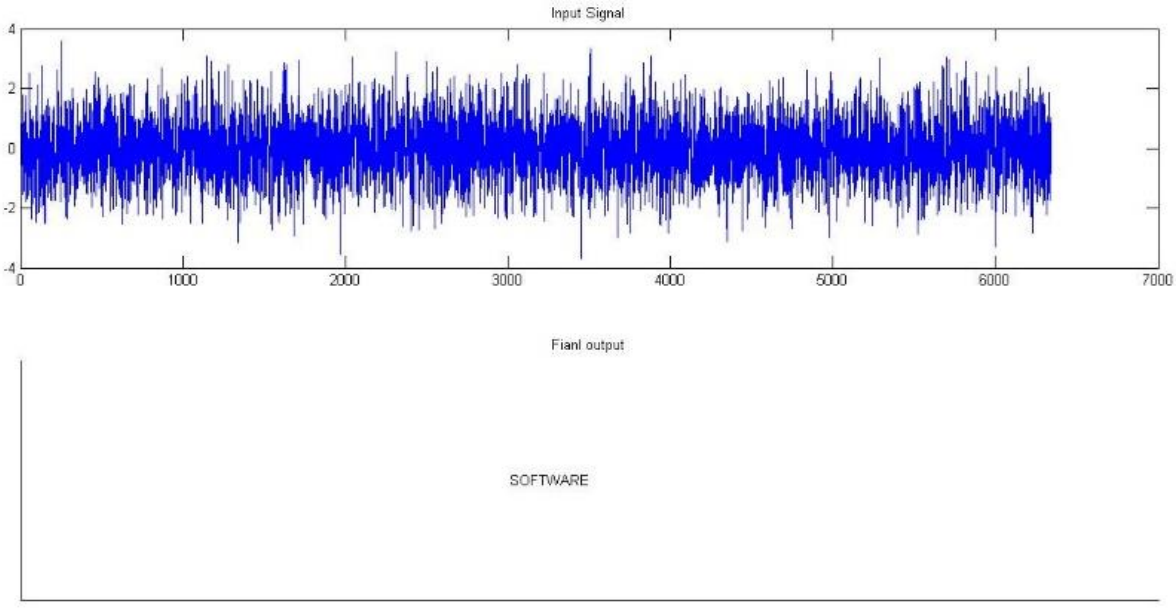


Figure 4.5: Final Output

The recognition accuracy performance with feature selection and without future selection algorithms are given in table 4.3.

Table 4.3: Recognition Accuracy

Method		Accuracy
Without Feature Selection		80%
With Feature Selection	No. of Iterations	
	100	81.65%
	200	83.7%
	300	90.21%

From table 4.3, it can be seen that an accuracy rate of 97.8% is obtained at300 iterations for optimized 187 features. From this, it can be concluded that the classification accuracy increased while using the optimal feature selection algorithm.

4.4 CHAPTER SUMMARY

This chapter presents a new and efficient scheme for automatic speech recognition. It works reasonably well for its easiness and lack of any more complex techniques; however, it is not super simple. Speech recognition using statistical and acoustic features and a support vector machine was implemented in this research. A significant issue in proposing a good Automatic speech recognition (ASR) system is the selection of a set of suitable input feature variables. So, a hybrid Artificial Bee Colony and Particle Swarm Optimization (ABC-PSO) is implemented for optimal feature extraction and selection of extracted features. ABC-PSO algorithm selects the most significant features among all features to increase the performance of the Automatic Speech Recognition system. The accuracy of the support vector machine classifier is used for selecting the most optimal set of features for the recognition of speech. Experimental results show that the proposed method can effectively recognize speech data.

CHAPTER 5.

FUZZY DWT BASED FEATURE SELECTION WITH CS-ANN CLASSIFIER

5.1 INTRODUCTION

Speech signals contain a gigantic measure of data and can be depicted as having various levels of data. Voice-based interfaces hold a way to understand this accomplishment. In this connection, Automatic speech recognition (ASR) systems in distinctive languages pick up significance. ASR is an imperative undertaking in digital signal processing related applications. It is the procedure of automatically changing over the spoken words into written text by the PC framework [134]. In recent decades speech recognition has made broad innovative advances in numerous fields, for example, call steering, automatic translations, data looking, data entry and so on. Speech recognition has been an expert by consolidating different algorithms drawn from diverse disciplines, for example, statistical pattern recognition, signal processing, semantics and so forth. The imperative uses of SR are security gadgets, household apparatuses, PDAs, ATMs, and PCs. The objective of the SR field is to create a method and framework to produce text from speech data to a machine. Given significant development in static displaying of speech, ASR nowadays gains far-reaching application in the tasks that need a man-machine interaction, for example, automatic processing of the call.

SR can be generally partitioned into two phases: extraction of features and classification. Even though noteworthy signs of progress have been made in SR innovation, however a troublesome issue to outline an SR framework for speaker-free, nonstop speech [135]. Among the largest, crucial inquiries is, in case the majority of the data important to recognize words is conserved while the phase of extraction of the feature. If imperative data is missing in the course of this phase, the execution of the accompanying classification phase is naturally handicapped and can under no circumstances match up to human ability.

Extraction of features can be comprehended as a stage to decrease the volume/largeness of the feed-in information, a diminishment that prompts some data loss. Normally, in SR, partitioning of the speech signals into segments is done and from each segment, features are extracted. In the process of extraction of features, audio signals are transformed into an arrangement of feature sets. At that point, these feature sets are exchanged for the

classification phase. For instance, in the case of dynamic time warping (DTW), this succession of feature vectors is contrasted with the reference information set. For the instance of hidden Markov models (HMM), vector quantization may be connected to the feature vectors, which can be seen as a further stride of feature extraction. In either case, data wasted in the process of moving from audio signals to a succession of vector sets must be maintained low.

In different many-sided applications, for example, speech recognition where the frameworks are created in light of genuine information, handling an extensive number of features is successive. Be that as it may, a significant number of the features are not pertinent to the issue of concern. Moreover, various features hold out for a lot of calculations, which back off the general procedure. Under these circumstances, naturally releasing insignificant features is important to accomplish a prototype that is exact and dependable in taking care of the issue at priority [136]. Also, utilizing the feature selection method lessens the computational expense, which improves the response of the procedure.

Feature selection (FS) is a standout amongst the most profitable investigation territories and has pulled in a lot of consideration in recent decades. For the FS undertaking, two surely understood strategies for feature assessment are utilized. One of them uses distance metrics to quantify the cover among distinctive classes. In this system, the probability thickness functions of the example appropriation can likewise be considered. Thus, the subset for which the normal overlay is insignificant is observed as an answer. In the Meantime, intra-and also between class distances can be estimated by giving the Entropy and fuzziness of the features. During that time system, classification errors in light of the feature subset candidates are assessed. Therefore, the subset with insignificant misclassification is chosen as an answer. A few techniques have been thought about and have been actualized for the variable selection. In the Meantime, utilizing flexibilities to characterize the optimization issue is useful [137]. For this reason, the fuzzy set theory is utilized to classify the adaptabilities for the objective functions. This strategy prompts accomplishing additional exchanges to tackle this issue. Fuzzy optimization systems have been now and again utilized as a part of the optimization of clashing objectives. In this study, an ASR system for the man-machine interaction with DWT based feature extraction and ANN classifier optimized with the Cuckoo search (CS) algorithm is developed.

5.2 DWT BASED FEATURE EXTRACTION AND CS-ANN CLASSIFIER

Considering the SR, the recognition of the speech signal is carried out by taking out more appropriate features that depict the signal. However, there are some approaches used for the extraction of features, the integration of AI in this arena will give improved outcomes. In this study, a procedure for man-machine interaction using discrete wavelet transform (DWT) based features along with a novel fuzzy-based feature selection method and Hybrid Cuckoo search – Artificial neural network (CS-ANN) classifier for automatic speech recognition system is suggested. The proposed method consists of three stages. They are Signal Pre-processing, Fuzzy based DWT and Classification. Initially, the input recorded speech signal is pre-treated for removing noise and detecting the word. Then, the pre-processed signal is subjected to a feature extraction process by using a discrete wavelet transform (WT) implemented to the speech signal where the pre-processed speech signal is disintegrated into many frequency channels using the characteristics of the WT. Here, in this proposed method, an 8-level multiresolution wavelet transform is employed, so that at each decomposed level eight types of features like Standard Deviation, Mean, Entropy, Skewness, Kurtosis, Log energy Entropy, Shannon entropy, and Renyi's Entropy of the speech signal are withdrawn. Since an 8-level DWT is used, so the number of features extracted is high. So, a Fuzzy model is employed to choose the optimum characteristics from speech signals which are extracted by DWT. Finally, the selected optimum set of features is used for training of ANN classifier and then depending on these features the spoken word is identified, and the resultant text will be shown. To improve the classification accuracy of the ANN the weights of the NNs are improved by employing the Cuckoo search (CS) algorithm. The suggested technique is applied in the MATLAB working stage and the results are matched with the previous methods under different conditions.

As presented in figure 5.1, the input speech signal is initially undergoing some sequence of preprocessing steps as sampling the signal, producing segments by Hamming windowing method and denoising the signal through the Harmonic decomposition process with the Harmonic decomposition method.

After preprocessing, the specified features from the signal are mined using DWT and from the extracted features only the optimum number of features is selected by the Fuzzy Inference System (FIS).

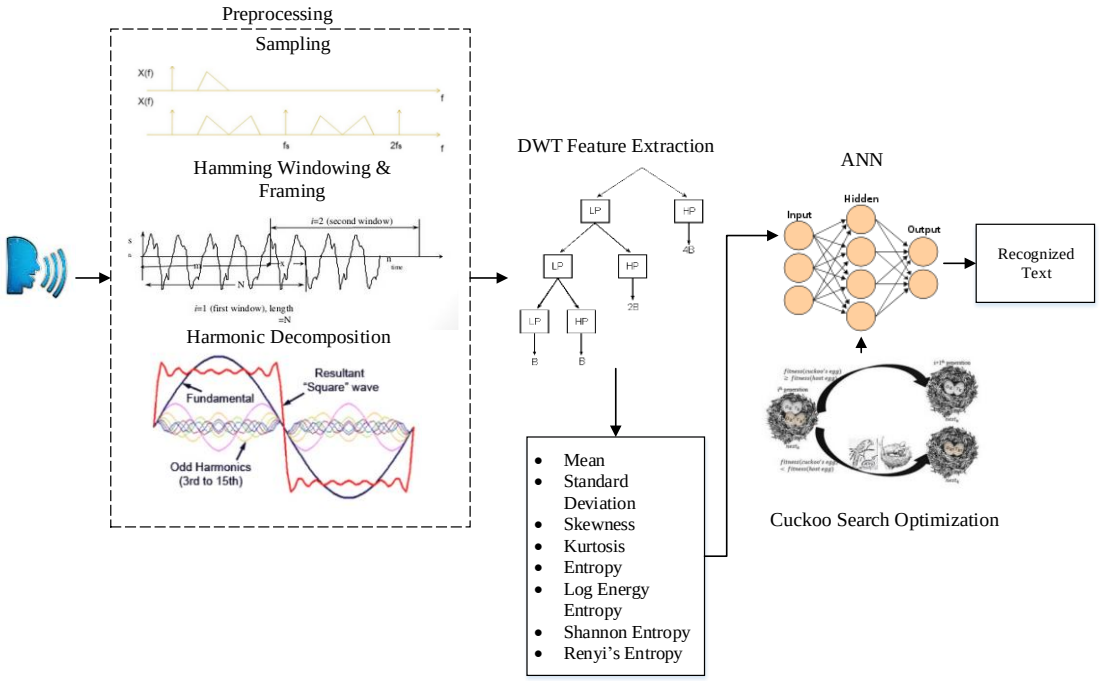


Figure 5.1: Schematic outline of the Proposed SR System

Each of these processes is explained in detail in the following sections. Consider that the database D contains the 'N' number of speech signals as mathematically represented as $D = \{s_1, s_2, \dots s_N\}$. From this, the signal $s_i, i = 1,2, \dots, N$ is taken out and further operations are applied as discussed in the following sections.

5.2.1 Pre-processing of Input Speech

The input speech signal produced by the human talk, normally processed using some pre-processing techniques to make the signal to be suitable for further operations being applied to the signal. The purpose of pre-processing the speech signal is to facilitate signal processing, eliminate the silent frequencies and remove the noises. Each of these pre-processing is performed through the following steps.

The sampling of the speech signal S_i is performed to facilitate further operations on the speech signal by converting it into a digital format. Several types of sampling methods or available, but for processing the speech signal a Bandpass Sampling method is employed in the proposed methodology.

For the bandpass signals, the spectrum of the signal *S(w)*will be zero for the range of frequencies except $f_1 < f < f_2$. In general, the frequency $f_1$of the bandpass signal is non-

negative and will be greater than zero also the aliasing effect is zero when $f_s < 2f_2$, where the sampling frequency is given as f_s, which is calculated using equation 5.1.

$$f_s = \frac{1}{T} \tag{5.1}$$

Where,$T = \frac{m}{2f_2}$ is the sampling interval and hence equation 5.1 is modified as in equation 5.2.

$$f_s = \frac{2f_2}{m},\ m < \frac{f_2}{B} \text{ and } f_s = \frac{2KB}{m},\ f_2 = KB \tag{5.2}$$

Where *B*= Bandwidth of the signal and *m*=Number of Replications used in the sampling process. It can be any integer till $f_s = 2B$. The sampled spectrum of the speech signal of bandwidth *B* and the minimum sampling rate f_s is given by equation 5.3.

$$S_s(\omega) = \frac{1}{T}\sum_{i=1}^{N}\sum_{n=-\infty}^{\infty} s_i(\omega - 2nB) \tag{5.3}$$

Where, *n* = time instant in discrete level, after sampling the points from the input speech signal, the samples are converted into frames by using the Hamming window and this is explained in the following process.

5.2.1.1 Framing and Windowing

The input signal produced may contain some silent sounds which are not necessary for the speech recognition system to recognize that signal. Hence the signal is converted into frames that smooth the progress of signal analysis. In the process of framing the signal is converted into frames with a period of 20-30 ms applied at 10 ms intervals with an overlap of 50% between adjacent frames and this is referred to as short-time spectral analysis.

Here, the Hamming window is employed, in the process of windowing which results in many frames of the speech signal. Hamming window is the famous windowing technique usually preferred in speech signal processing as it has its advantages. The process of windowing with Hamming window on speech signal is explained as follows:

In figure 5.2, the L point Hamming window in the time and frequency domain is displayed where L is the length of the window, and it is considered here as 64. The Hamming window is generally represented as given in equation 5.4 [26].

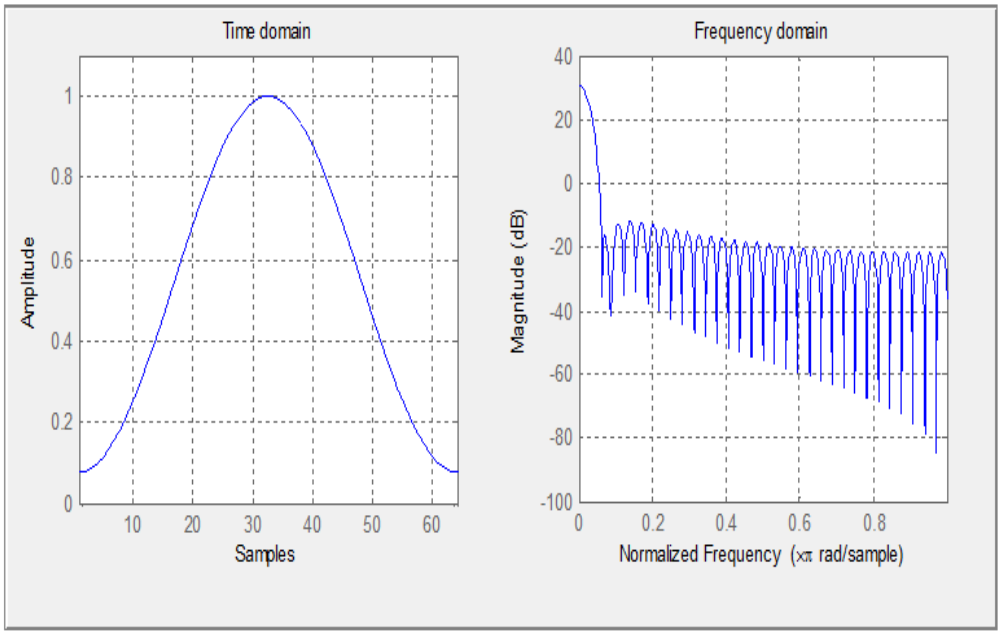


Figure 5.2: L point Hamming Window in Time and Frequency Domain (L=64)

$$w(n) = 0.54 - 0.46(\frac{2\pi(n-1)}{N}) \quad (5.4)$$

Where, *N*= Number of samples in each frame which is equal to L, L-1, or L+1.

In the process of windowing and framing, the sampled signal is initially divided into frames with the specified time intervals and then the hamming window is applied to each frame to generate a signal which is free from discontinuities.

Consider that the feed-in speech sample is represented as $s_i(n)$ for the i^{th} sample, then the resulting frames after being processed by the hamming window are given by the equation 5.5.

$$y_i(n) = \sum_{i=1}^{S} w(n)s_i(n) \quad (5.5)$$

Where *S* = Total number of frames produced

$s_i(n)$ = nth sample from the spectrum $S_s(\omega)$

The resulting *S* number of frames produced by equation 5.5 is the speech frequencies that are free from discontinuities. These frames are further enhanced by the reduction of the

noise present, and this is performed by the Harmonic decomposition process as explained in the following section.

5.2.1.2 De-noising frames by Harmonic decomposition

After converting the speech samples into frames without discontinuities the noise present in these samples is removed using the denoising method and the type of noise present in the speech signal is the harmonic noise. Hence, the Harmonic decomposition method is employed here as the noise reduction technique in which the input speech sample is decomposed into its fundamental as well as harmonic components.

After, decomposing the speech sample into components the order of the components which are greater than three could be ignored.

The speech sample obtained from the windowing function is then denoised$y_i(n)$ by correlating it with the fundamental and harmonic components up to the order of three which is given mathematically in equation 5.6.

$$y_{di}(n) = \sum_{i=1}^{S} \sum_{j=1}^{3} (y_i(n) * h_j) \tag{5.6}$$

Where, $y_{di}(n)$ = Denoised sample, and
h_j= Harmonic Levels of the input speech sample.

Thus, the correlated signal produced is free from harmonic noise by filtering out the harmonic component after reconstruction. The pre-processed speech samples are then fed to the feature extraction stage, in which the features for the speech signal aid the classification in the recognition phase. The feature extraction phase is given in the following section.

5.2.2 Fuzzy Based DWT Feature Extraction

After, producing the de-noising signal of the speech input the features are extracted from the pre-processed signal to recognize the speech by training the classifier in the classification stage.

There are several methods employed to withdraw the features from the speech signal, but the incorporation of artificial intelligence could produce the most suitable features for the recognition of the speech signal. Hence in the proposed methodology, a novel fuzzy-based

DWT feature extraction is employed where eight different types of features are extracted such as Mean, Standard Deviation, Skewness, Kurtosis, Entropy, Shannon entropy, Log energy Entropy and Renyi's Entropy from the signal using DWT feature extraction with eight-level decompositions.

After the features are extracted most of them are not having the necessary details to recognize the signal so the fuzzy model is used to select the most suitable features. The process of DWT of extracting the features and the feature selection through the fuzzy model is detailed as follows.

5.2.2.1. *DWT in Feature Extraction*

A linear conversion that works on a data vector is a discrete wavelet transform (DWT) whose dimension is an integer power of two, converting it into a statistically diverse vector of identical size. In image compression and signal processing, the wavelet transform has obtained extensive acceptance. A signal's multi-resolution depiction is offered by the DWT which is very beneficial in examining "real-world" signals and therefore this transformation is accepted to excerpt the speech signal features in the proposed approach.

Basically, by a DWT a distinct multi-resolution depiction of a time persistent signal is acquired. It transforms a sequence $a_0, a_1, \dots a_n$ into one lowpass coefficient series known as "approximation" and one high pass coefficient series known as "detail". '*n/2*' is the size of every sequence. In real-life conditions, such conversion is implemented repeatedly on the low-pass series until the preferred number of repetitions is attained.

The function is not incessant and henceforth not distinguishable. Daubechies wavelets are the families of wavelets whose inverse wavelet transforms are adjoint of the wavelet transform i.e., they are orthogonal. Using Daubechies wavelets the wavelet transform gives rise to gradually finer discrete samplings making use of repetition relationships.

Each resolution scale is two-fold that of the earlier scale. From data compression range to signal coding, the discrete wavelet transform has a wide range of applications. Hence, it is a device that divides information into varied frequency modules, and afterwards, every single component with a resolution according to its scale is considered. DWT is calculated with a cascade of filters ensured by a factor of two sub-sampling.

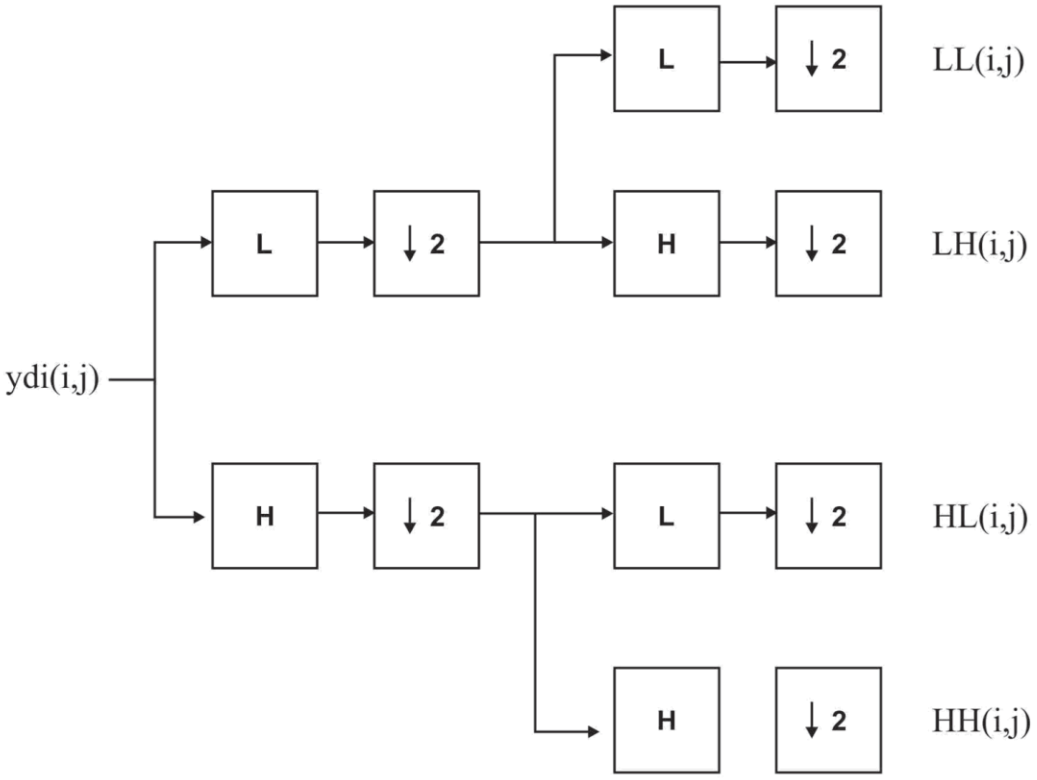


Figure 5.3: General architecture of the 1-level DWT

In figure 5.3 the fundamental DWT architecture is shown. Here, one level DWT is shown in which the speech signal is first split into low and high-frequency bands with the Low Pass and High Pass filters as given by the impulse responses denoted by *g(m)* and *h(m)* respectively and this is given in equations 5.7 and 5.8 and they are divided further to produce two low-frequency bands and two high-frequency bands, and this is the complete one level decomposition by DWT. The ↓ symbol denotes the downsampling of the filtered element.

$$l_{i+1}(m) = \sum_{n=-\infty}^{\infty} y_{di}(n-2m)g(m) \tag{5.7}$$

$$h_{i+1}(m) = \sum_{n=-\infty}^{\infty} y_{di}(n-2m)h(m) \tag{5.8}$$

The process is given in equations (5.7) and (5.8) and can be continued until the required level of decomposition is reached by extending the architecture as given in figure 5.3. The DWT's major feature is a function of multiscale representation. At different levels of resolution given function can be examined using the wavelets. The DWT is also orthogonal and could be invertible. In this proposed methodology, an 8-level DWT is employed and

hence the filtered coefficients produced for each signal is 32. From these coefficients, the features mentioned above are extracted in the following manner.

(i) Mean, μ : The Mean value of the filtered coefficients is calculated at each level of the decomposition. If the low and high-frequency components of level *k* are represented as l_{lk}, l_{hk}, h_{lk} and h_{hk} respectively, then the Mean value is calculated as given in equation 5.9.

$$Mean, \mu = \frac{1}{n}\sum l_{ik}\, h_{ik}, \qquad i = l \;\&\; h;\ n=4 \tag{5.9}$$

(ii) Standard Deviation, σ:

The Standard Deviation of the calculated filtered coefficients can be measured from the variance of the coefficients, and it is the estimate of how the sample in the signal is far from the Mean point. The variance and the Standard Deviation from that are calculated as given in equations 5.10 and 5.11.

$$Variance,\ \sigma^2 = \frac{1}{n}\sum (x_i - \mu) \tag{5.10}$$

$$Standard\ deviation,\ \sigma = \sqrt{\frac{1}{n}\sum (x_i - \mu)} \tag{5.11}$$

(iii) Skewness, S:

A significant distribution parameter is symmetry. Based on the definition, a skewed variable Mean is not situated at the distribution centre. It might be observed that two signal segments can have the same Mean and Standard Deviation but diverse values for Skewness. Skewness is conveyed as in equation 5.12.

$$Skewness, s = \frac{E(x_i - \mu)^3}{\sigma^3} \tag{5.12}$$

The positive and negative values of 'S' specify that the signal is skewed right or left, correspondingly.

(iv) Kurtosis, k :

The extreme distribution degree is offered by Kurtosis and conveyed as in equation 5.13.

$$Kurtosis, k = \frac{E(x_i - \mu)^4}{\sigma^4} \tag{5.13}$$

(v) Entropy, $H(s_n)$:

On the DWT application signal, a discrete wavelet coefficients series D_i under different scales *j (j = 1, 2..., m)* are attained, which can reflect the time-frequency distribution to some extent. By information theory, the signal entropy is computed from these filtered coefficients and the Entropy calculation is provided as follows. Any event uncertainty is related to its probabilities and states. All possible states aggregation is known as sample space$\{S_1, S_2, \ldots . S_n\}$. Each bit of information has a probability $P(S_n) = P_n, 0 \leq P_n \leq 1, \sum P_n = 1$. Then the information quantity of the samples obtained from the wavelet is calculated as in equation 5.14.

$$I(S_n) = -logP_n \tag{5.14}$$

$I(S_n)$ is a random variable varying with diverse information, so it is not apt for the entire information source measurement. Hence, the mathematical expectation of the self-information is defined as the Mean self-information of the information source, which is the Entropy signified by $H(S_n)$ and given in equation 5.15.

$$\mathrm{H_{sh}(S_n)} = \mathrm{E[I}(S_n)] = -\sum P_n logP_n \tag{5.15}$$

(vi) Shannon Entropy $\{\boldsymbol{H_{sh}(S_n)}\}$**:**

Shannon introduced the Shannon Entropy which concludes the predictive value of the comprised information in a message, in the communication field. In principle, Entropy is a measure of signal uncertainty. The calculation of Shannon Entropy is given in equation 5.16.

$$H_{sh}(S_n) = \frac{H(S_n)}{\log m} \tag{5.16}$$

Here, m specifies the number of bins on the divided signal.

(vii) Log energy Entropy $\{\boldsymbol{H_{log}(S_n)}\}$**:**

The Log energy Entropy is given by the samples is calculated as given in equation 5.17.

$$\mathrm{H_{log}(S_n)} = -\sum(\log(P_n))^2 \; H_{\log}(s_n) = -\sum\left(\log(P_n)\right)^2 \tag{5.17}$$

(viii) Renyi's Entropy {$H_{ren}(s_n)$ }:

The simplification of the normal model of Entropy is Renyi's Entropy and in fact, Renyi's Entropy is adjacent associated with free energy and depends on the variable $\alpha \geq 0$. For the filter coefficients measured from the speech signal, this Entropy is calculated as in equation 5.18.

$$H_{ren}(s_n) = \frac{1}{1-\alpha}\sum(\log(P_n)^{\alpha}) \tag{5.18}$$

The features given in equations 5.9 to 5.18 are calculated at each level of the decomposition and then the optimum features in each level are selected using the fuzzy model and the feature selection process is explained as follows.

5.2.2.2. *Fuzzy Based Feature Selection*

The wavelet features extracted as above are larger in number and such a large quantity is not necessary to recognize the input signal. For that purpose, the feature selection process is usually conducted after the feature extraction phase and the familiar feature reduction techniques used are Principal Component Analysis (PCA), Isomap techniques, etc. But these techniques consume more time and result in considerable information loss. Hence including human intelligence can produce apt features that are used better for the recognition of the speech signal. In this proposed methodology the required features employing fuzzy logic are selected. Here, the method engaged is the fuzzy feature estimation index for a set of features which is described in terms of membership values indicating the degree of similarity between two features.

In fuzzy-based feature selection, the parameter called evaluation index is calculated between a set of features based on the calculated membership functions. Initially, the degree of membership $\mu_{uv}$is measured between the feature sets *u* and *v*. With that membership function the evaluation index is measured and the feature sets which are having a minimum value of the index are selected in training the neural network. The membership function μ_{uv} for the features is calculated as given in equation 5.19.

$$\mu_{uv} = \begin{cases} \left(1 - \frac{d_{uv}}{D}\right), & if \;\; d_{uv} \leq D \\ 0, & otherwise \end{cases} \tag{5.20}$$

In equation 5.20, d_{uv} is the Euclidean distance which measures the similarity between the two features and *D* is the minimum distance between any given two features. The distance d_{uv} is calculated as in equation 5.21.

$$d_{uv} = \sqrt{(u - v)^{\,2}} \tag{5.21}$$

And finally, the evaluation index of the membership function is measured for all the features in the feature space as in equation 5.22.

$$e = \frac{1}{s(s-1)} \sum_u \sum_v \mu_{uv} \, (1 - \mu_{uv}) \tag{5.22}$$

Where 's' is the no. of feature samples present in the respective feature space. This fuzzy evaluation index is determined for all the eight features and then finally the respective features which are having a minimum value of the index are considered in the further recognition process. After selecting the optimum features from the feature space, the training and the testing of the ANN are carried out in the next stage in which the optimization of the network is done utilizing the Cuckoo search algorithm and the process involved in this stage is detailed in the following section.

5.2.3 Recognition of Speech with CS-ANN Classifier

The optimally selected features from the feature extraction are then used for training the ANN to recognize the speech signal, but in normal backpropagation algorithm is used for the optimization of the weight function between the links. But here the CSO algorithm to produce fair optimization of the network is employed. The steps of this training procedure are explained as follows. But before that, the basic introduction to ANN is given.

5.2.3.1 Artificial neural network (ANN)

A distinctive ANN has two types of basic constituents. They are neurons that are processed elements and links which are interconnections between neurons. Each link has a weighting parameter. From other neurons, each neuron gets a stimulus, progresses the information, and an output is produced. Neurons are classified into input-output (I/O) and hidden neurons. The I/O layers are called the first and the last layers correspondingly, and the

enduring layers are called hidden layers. Consider in the k^{th} layer the no. of neurons n. Let, the weight of the link is represented as w_{ij}^{k} between the j^{th} neuron of the $(k\text{-}1)^{th}$ layer and the i^{th} neuron of the k^{th} layer. If each neuron $w_{i0}^{k}$is an additional weighting parameter, signifying the bias in the i^{th} neuron of the k^{th} layer. Before the neural network training, the weighting parameters are initialized. Systematically they are iteratively updated during training. Once the neural network is finished, the weighting parameters remain stable throughout the neural network usage as a model.

The training process of an ANN is to modify the biases and weights. The most prevalent technique to train, feed forward ANNs in numerous domains is Back Propagation (BP) learning. Though, one drawback of this method, which is a gradient-descent method, is that it necessitates a differentiable neuron transfer function. Moreover, as NNs create intricate fault surfaces with many local minima, instead of a global minimum the BP inclines to converge into local minima. In the modern era, several enhanced learning algorithms have been suggested to overwhelm the handicaps of gradient-based methods. Because of conventional numerical methods' computational shortcomings in resolving complex optimization problems, researchers may depend on meta-heuristic algorithms. Over the last years, numerous meta-heuristic algorithms have been applied successfully to several engineering optimization problems. For many complex real-world optimization problems, better solutions have been delivered by them in comparison with conventional numerical methods. Here, Cuckoo Search optimization as a metaheuristic algorithm for optimizing the ANN is employed. The reason for this is explained in the consecutive section.

As shown in figure 5.4 below, the input features are fed with the input layer at each input neuron. With this input, the basis function is calculated at the hidden neurons present in the hidden layer with a weighting function assigned to each link between the input layer and the hidden layer. Then the basis function is calculated at each output neuron of the output layer separately. These steps are repeated with another sample of the same words present in the signal and if any error is produced, means the weighting function is optimized using the CSO algorithm. The training and testing procedure of the ANN in recognizing the spoken words is explained as follows and the kind of Neural Network used here is the Feed Forward Neural Network (FFNN).

The training of ANN is explained in the following steps:

Step 1. Define the selected features from the fuzzy evaluation index as the input neurons. Eight types of features in recognizing the word and the fuzzy-based feature selection technique produces the two best feature values for each type.

Therefore, 16 feature values will be produced for training the NN as the input neurons.

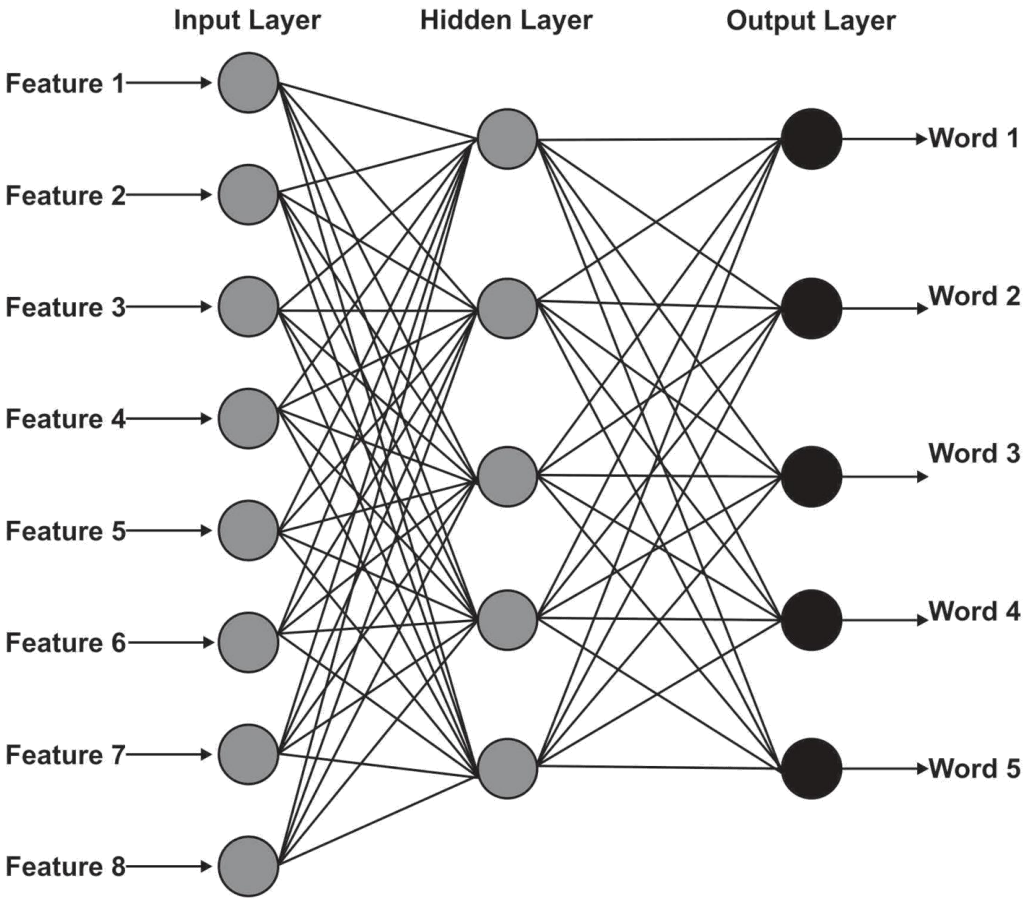


Figure 5.4: Basic Architecture of an ANN

Step 2. Then the basis function at each input neuron is to be calculated for each of the hidden neurons with the weighting function. This basis function calculation is given by equation 5.23.

$$I_b = \sum_{j=1}^{J} u_j \, w_{jk}^{i} \tag{5.23}$$

Where J = Number of input neurons

u_j = Value of feature at each input neuron

w_{jk}^{i} = weight value of the link between input neuron j and hidden neuron k

Step 3. After calculating the basis functions at the input neurons, the activation function is calculated from that at each hidden neuron as given by equation 5.24.

$$A_k = \left(1 + exp(-I_b)\right)^{-1} \tag{5.24}$$

Step 4. Once the activation function is measured by each hidden neuron Means, the basic function of each of the output neurons is determined. This basis function calculation at the output side is given in equation 5.25.

$$O_b = \sum_{k=1}^{K} A_k w_k^o \tag{5.25}$$

Where w_k^o = Weight value between the links of the hidden neuron and output neuron. The basis function produced at the output side is the expected output to be produced at the recognizer. Sometimes, the value at the output neuron may deviate from the actual output and this is called learning error and denoted by the equation 5.26.

$$Learning\ Error, E_L = \frac{1}{2}(O_{act} - O_{obt})^2 \tag{5.26}$$

Where, O_{act} = Actual Output and, O_{obt} = Obtained Output.

Step 5. Optimization of Weight coefficients: The Cuckoo search optimization is employed in this part to select the corresponding weight coefficients so that the produced learning error is zero. Hence the fitness function used here is the learning error constrained to the weighting functions. The steps employed in the CSO algorithm are given in the following part.

5.2.3.2 Cuckoo search Optimization in Output, Weight Updating

Yang and Deb developed the Cuckoo search algorithm in 2009. It is an important modern optimization algorithm that replicates some cuckoo species' breeding performance. Modern investigations have exposed that Cuckoo Search is possibly much more effective as compared to PSO and GA. So that the CS algorithm is employed here for the optimization of the neural network to produce the zero-learning error. The nature of the cuckoo search algorithm is detailed as follows.

The Cuckoo proliferation approach is used. In communal nests, some cuckoo species lay their eggs, though moderately several species employ obligate brood parasitism by laying

their eggs in the host birds' nests (frequently other species). Brood parasitism principally undertakes three classes that is intra-specific brood parasitism, cooperative breeding, and the nest take over. Once the eggs are oviposited, if the host birds could notice or find out that the eggs are not their own, they would possibly smash the unknown eggs or leave their nests and build new nests elsewhere; however certain female cuckoo species can oviposit very specified in imitation in the shape and design as of the host bird's eggs. This reduces the chances of revealing their eggs.

Many pieces of research have presented that a lot of animals and insects' flight behaviour might comply with specific distinctive individualities of levy flights. A common concern of levy flights and random walks to attain a new solution is offered in equations 5.27 and 5.28.

$$x^{i+1} = x^i + \beta \oplus Levy(\lambda),\ \beta > 0 \tag{5.27}$$

$$Levy(\lambda) = t^{-\lambda},\ 1 \leq \lambda \leq 3 \tag{5.28}$$

Where, x^{i+1} = New solutions (Here, learning rate). By random levy walk, some of the new solutions must be produced nearby the finest resolution. Though, a significant fraction of new solutions must be formed by far-field randomization. This would assure the algorithm is not stuck in a local ideal situation.

To design the regular CS algorithm, the following 3 ideal guidelines are made.

- Only a single egg at a particular time is laid by each cuckoo, and in an arbitrarily chosen nest, it is dumped.
- The finest nests with great quality eggs (solutions) would be transferred to the subsequent generation (algorithm iteration).
- The available no. of host nests is fixed, and the probability of discovery of each cuckoo's egg by the host bird is given as $P_a \in [0, 1]$.

Conferring to the abovementioned 3 rules, the fundamental stages of Cuckoo Search can be precise as the pseudo-code signified in figure 5.5. Initially, the population (host nests) is defined with the weight coefficients as given in the above section. Then as given in the flowchart depicted in figure 5.5 below, the fitness function is calculated here. The fitness function is the learning error of the Neural Network, and the objective is the reduction of the error with the best weighting coefficients. Once the fitness is calculated means the new

host nests are formed through the levy flights by considering nest '*i*' and the fitness is calculated again.

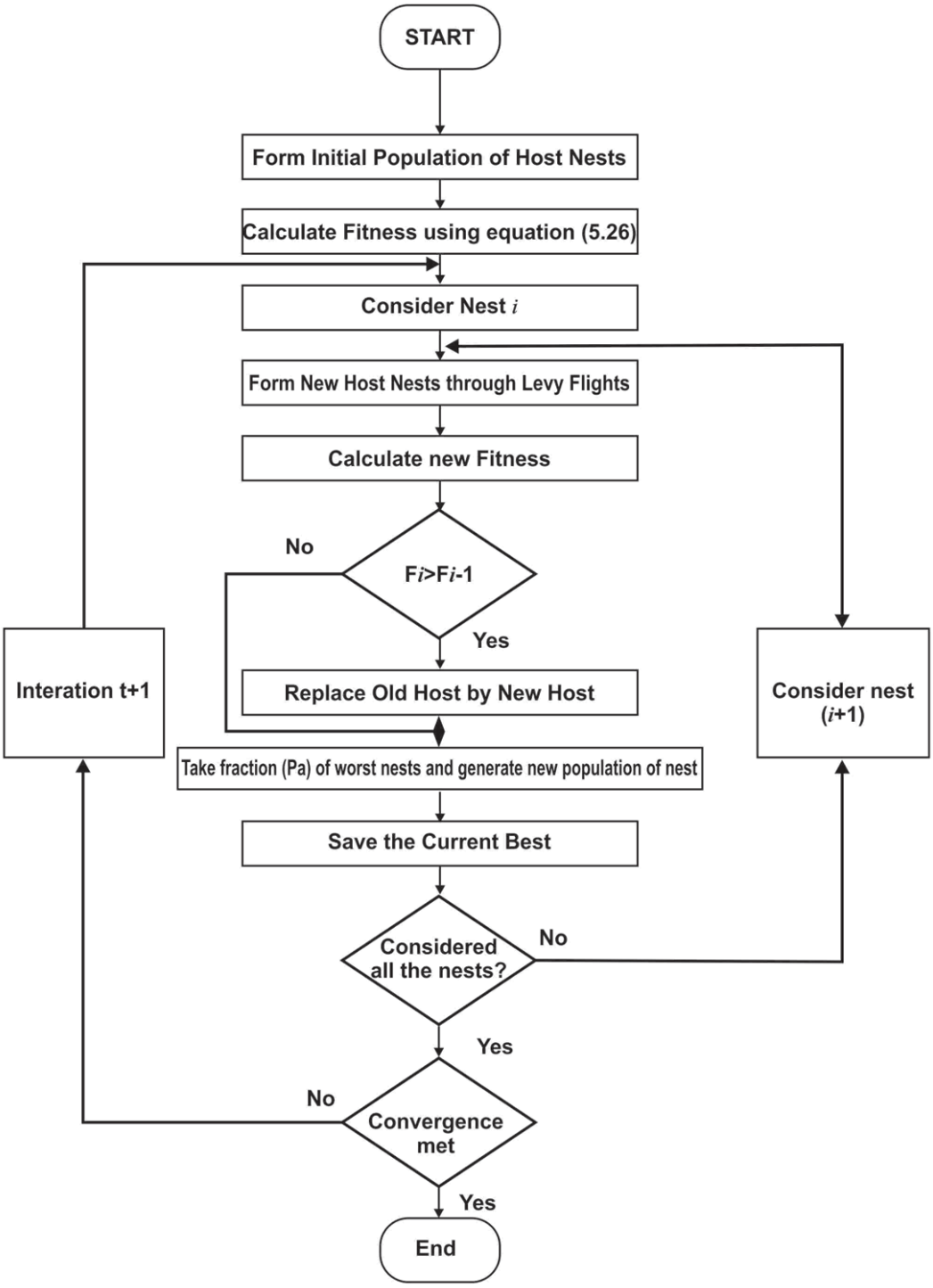


Figure 5.5: Cuckoo Search Algorithm

Then the calculated fitness value is compared with the one which is obtained in the previous stage and if the current one is better, the old host nest will be replaced with the newest one.

Else the fraction of the worst nests are taken to form the new population of nests. The best solution is obtained until the process will be stored and if all the nests are considered or the maximal count of iterations is attained means the algorithm will be stopped and the best solution will be returned. Otherwise, it will continue until are the nests are considered or else the maximal count of iterations is attained. The weight coefficients are optimized in this manner and the network is trained to produce the corresponding output values. After training the network with the maximum number of training samples the testing of the network is carried out and the testing procedure involved here is explained in the next section.

iii) Testing of ANN

The testing procedure is also like that of the training of the neural network except that the learning error and hence the optimization is not done here. In the training stage, the NN is trained to produce the words as recognized in the speech signal and in testing, the recognition performance of the network is validated. The experimental set-up employed in the proposed work, the results, comparison with other works and the corresponding discussions are given in the investigational part.

5.3 RESULTS AND DISCUSSION

In this thesis, an approach of the ASR system for the man-machine interaction with fuzzy based DWT feature extraction and the ANN optimized with the CSO algorithm is suggested. The proposed work is realised in the functioning MATLAB platform of version R2013a with the system configurations of the Intel Core i3 processor, 4GB RAM, and Windows 8 Operating system. In this section, the dataset used in the proposed work, the results of the suggested method, the comparison results as well as the discussions about the improvement of work are presented.

5.3.1 Dataset

In this thesis, the Grid corpus as a small-vocabulary ASR assignment to assess every approach is utilized. The Grid database contains 34,000 sentences that were spoken by thirty-four persons, i.e., 1000 sentences per person. The job of the Grid corpus is to identify sentences from a small vocabulary (51 words) with a fixed grammar of the form: command-colour-preposition letter-digit-adverb. A hundred speech data from the database is taken and among them, 80% is used for training the recognizer and 20% is for testing the

recognizer. The structure of the sentences in the grid database is given in the following table 5.1.

Table 5.1: Sentence Structure in Grid database

Command	Color	Preposition	Letter	Digit	Adverb
Bin	Blue	At	a-z (Except 'w')	0-9	Again
Lay	Green	By			Now
Place	Red	On			Please
Set	White	With			Soon

5.3.2 Preprocessing Results of the Speech Signal

From this signal, the preprocessing steps are applied to produce the enhanced version and to be suitable for the application of further operations. Initially, sampling of the signal is done to produce the sampled data at a sampling frequency of f_s = 10 kHz. After sampling, frames are produced with Hamming windowing at the specified intervals and then noise present in the signals is removed with Harmonic Level decomposition. The preprocessing result of the sample signal from the database is presented in figure 5.6.

In this figure, the top left corner shows the original speech signal, the sampled signal produced is given in the top right corner, the frames of the signal are given in the bottom left corner and the filtered signal is shown in the bottom right corner. After those eight different feature coefficients are extracted through 8-level DWT, which are 64 coefficients for each signal. From these two feature coefficients are selected optimally per each type of feature using the fuzzy model. The decomposed signal obtained with DWT feature extraction is depicted below in figure 5.7.

The eight types of features calculated from these decomposition levels for the speech signal ‘s’ are tabulated in the following table 5.2. Then the selected features from table 5.2 are used to train the NN in which the learning error is corrected by optimizing the weights at the output neuron through the CSO algorithm. The outcomes of speech recognition are shown in the next section. Figure 5.6 shows the preprocessing of a sample speech signal. The input speech signal, the framed signal and the filtered signal are shown here.

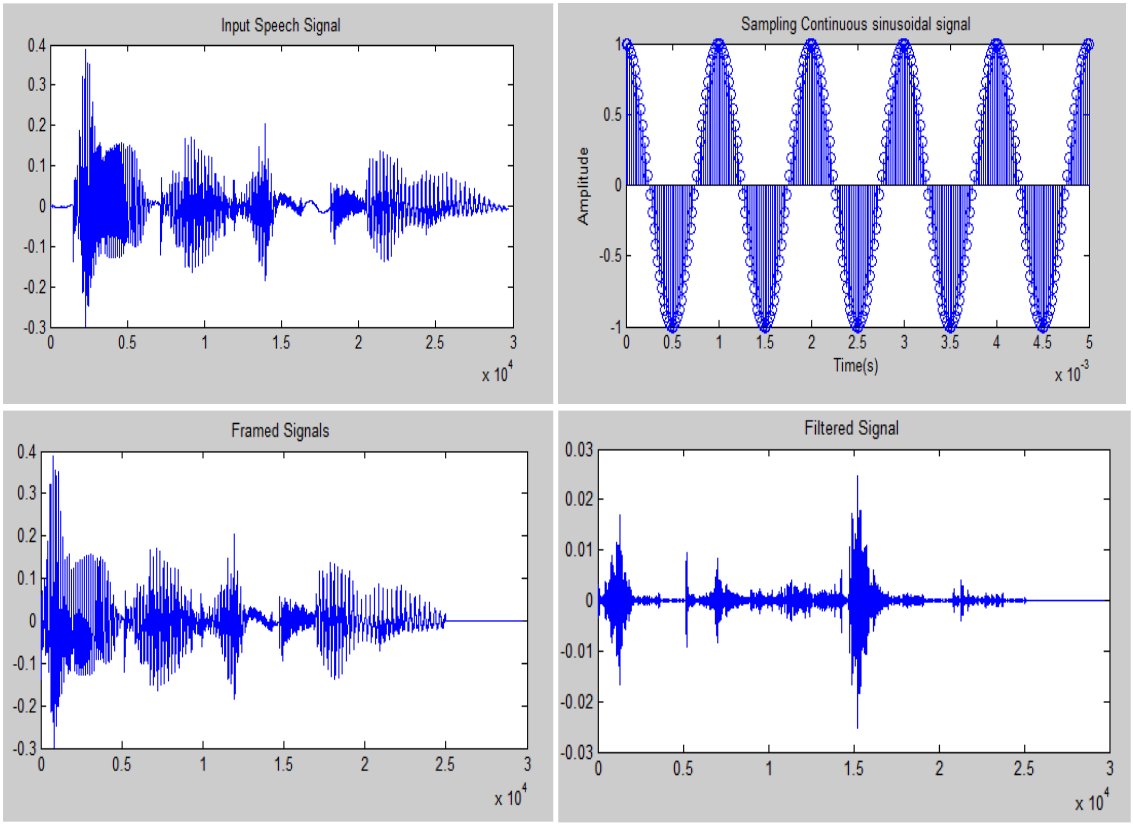


Figure 5.6 Preprocessing result of the sample speech signal

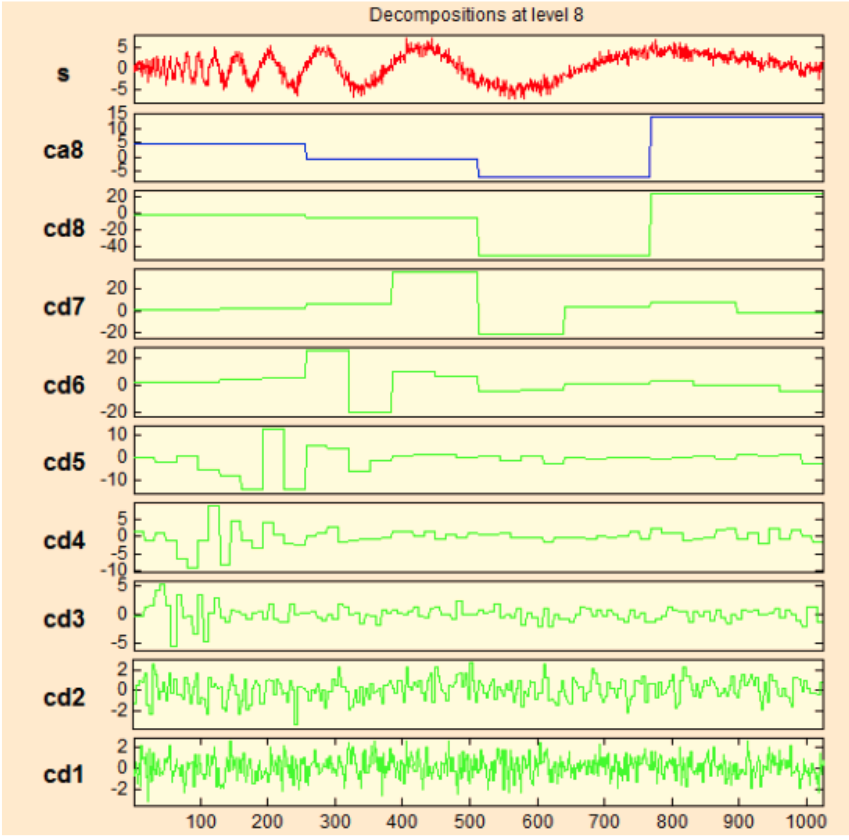


Figure 5.7: Eight level decomposition of the original speech signal by DWT

Table 5.2: Feature values of the signal S_i by DWT

Type of Feature	Feature Value
Mean	65.6667
Standard Deviation	58.0108
Skewness	-2.7095
Kurtosis	6
Entropy	0.5441
Shannon Entropy	-671.1688
Log energy Entropy	99.9938
Renyi's Entropy	0.00000078

5.3.3 Results of Recognition

In this section, the results of the proposed speech recognition system for man-machine interaction are presented. The working of the proposed methodology is assessed here depending on the performance metrics such as recognition accuracy, word error rate, sensitivity, specificity, true positive rate, and false-positive rate.

(i) Recognition Accuracy

Recognition accuracy is defined as the performance metric which is used to measure the performance of a recognition system. The recognition accuracy, R_a is simply calculated using the following equation 5.29.

$$R_a = \frac{Number\ of\ recognized\ words}{Total\ number\ of\ words} \tag{5.29}$$

The greater the value of R_a, the greater its recognition performance.

(ii) Word Error Rate (WER)

Word error rate is also the working measure used to measure the recognition system performance. This is calculated directly from equation 5.29 and given in equation 5.30.

$$WER = 1 - R_a \tag{5.30}$$

The lower the WER the better the performance of the speech recognition system.

(iii) Sensitivity

A measure of the capability of a system to properly identify positive samples is sensitivity. It could be computed using the following equation.

$$Sensitivity = \frac{TP}{TP+FN} \tag{5.31}$$

The sensitivity value ranges between 0 and 1, where 1and 0 mean best and worst recognition of positive samples, correspondingly.

(iv) Specificity

A measure of the capability of a system to identify properly negative samples is specificity. It might be computed using the following equation.

$$Specificity = \frac{TN}{TN+FP} \tag{5.32}$$

The specificity value ranges between 0 and 1, where 0and 1 refer to the worst and best appreciation of negative samples, correspondingly.

(v) False Positive Rate (FPR)

It is the existing rate of positive test results in matters that do not know to have the behaviour for which an individual is being tested. The FPR is computed as in the following equation 5.33.

$$FPR = \frac{FP}{FP+TN} \tag{5.33}$$

(vi) False Negative Rate (FNR)

It is the occurrence rate of negative test results in subjects referred to have the performance for which an individual is being verified. The FNR is computed as in the following equation 5.34.

$$FNR = 1 - TPR \tag{5.34}$$

(vii) ROC Curve

ROC curve depicts the characteristic of any of the recognition systems and it is the curve drawn between the TPR and FPR of the system. The ROC of the proposed methodology is shown in figure 5.10. The results of the proposed methodology in terms of these performance metrics are given in table 5.3 as well as in Figures 5.8 and 5.9.

The results of the proposed methodology in terms of recognition accuracy and the word error rate are represented graphically in figure 5.8. Table 5.3 gives the values of the performance metrics of the proposed fuzzy-based selection of DWT features with Cuckoo search Artificial Neural network Classifier.

Table 5.3: Results of Proposed Methodology

Performance Metric	**Result**
Recognition Accuracy (%)	95%
Word Error Rate (%)	5%
Sensitivity	0.95
Specificity	0.5
False Positive Rate (FPR)	0.5
False Negative Rate (FNR)	0.05

The performance of the recognizer is more understood by the performance metrics such as sensitivity, specificity, FPR and FNR for the results obtained with the methodology are

0.95, 0.5, 0.5 and 0.05 respectively. This shows that most of the words are correctly recognized by the proposed methodology for each signal. Also, the recognition accuracy of 95% shows a very good performance of the system. The word error rate of 5% is very less in this fuzzy-based selection of DWT features with the Cuckoo Search Artificial Neural Network Classifier proposed methodology.

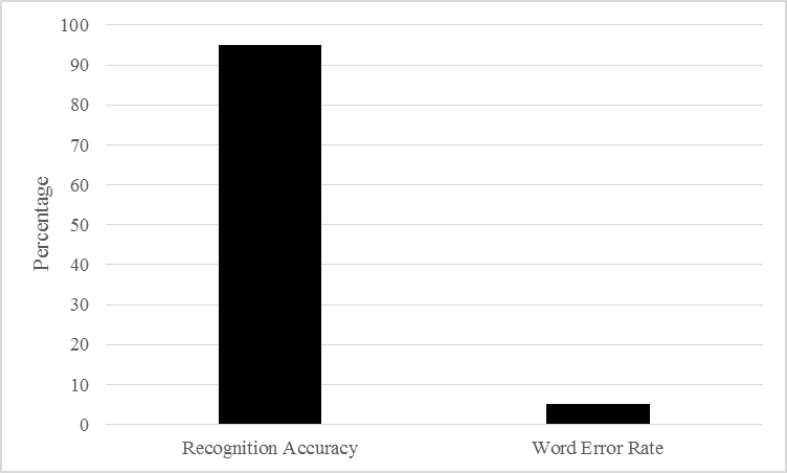


Figure 5.8: Performance of proposed methodology in terms of recognition accuracy and word error rate

The results of the proposed method in terms of these parameters are presented in the following figure 5.9.

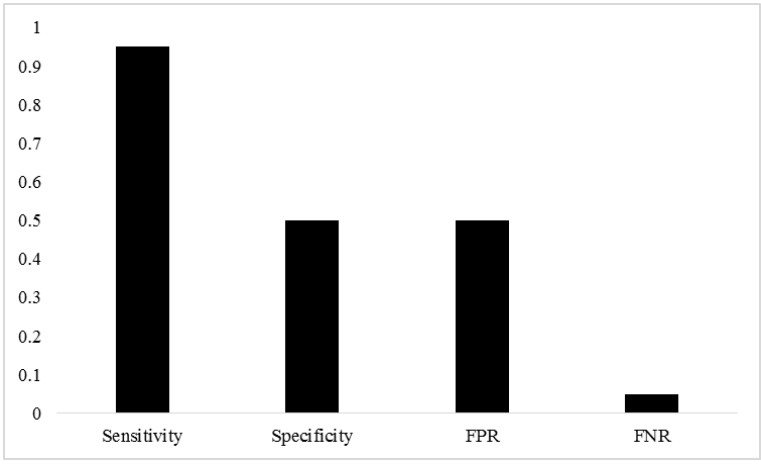


Figure 5.9: Performance of proposed methodology in terms of Sensitivity, Specificity, False Positive Rate (FPR), False Negative Rate (FNR)

The corresponding ROC graph is presented in the following figure 5.10. It is the graph drawn between TPR (Sensitivity) and FPR (1-Specificity).

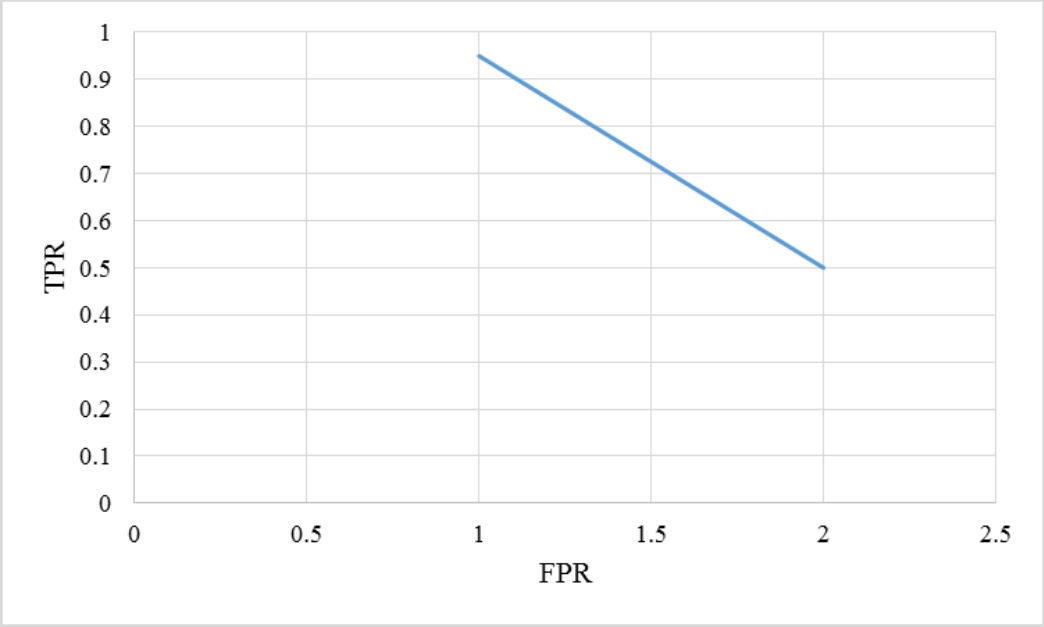


Figure 5.10: ROC curve of the proposed methodology

To demonstrate the effectiveness of the suggested methodology, it is matched with different techniques in terms of the performance metrics, and this is presented in the following section.

5.3.4 Performance Comparison

The performance of the proposed fuzzy-based selection of DWT features with the Cuckoo Search ANN Classifier used for the man-machine interaction system is validated. The results are compared with conventional backpropagation speech recognition and uncertainty-decoding based Automatic Speech Recognition. Under the presence of babble noise at three different levels of 0 dB, 5 dB and 10 dB, the performance of the three systems is compared and presented in tabular form. Also, a minimum of three and a maximum of eight features are used to compare the recognition accuracy and the word error rate. The performance metrics are evaluated and shown in table 5.4. The comparison result is shown in graphical form in figure 5.11 for ease of comparison. It shows the performance Comparison of the Proposed Methodology of ASR with Cuckoo Search ANN Classifier, ASR without Cuckoo Search ANN (conventional Back Propagation ANN) and Uncertainty decoding-based ASR.

Table 5.4: Comparison of Proposed Methodology with other techniques

Performance Metric	Method														
	Proposed ASR with CS-ANN						ASR without CS-ANN						Uncertainty Decoding based ASR (UD-ASR)		
	With 3 features (SNR in DB)			With 8 features (SNR in DB)			With 3 features (SNR in DB)			With 8 features (SNR in DB)			UD-GHF Decoding (SNR in DB)		
	0	5	10	0	5	10	0	5	10	0	5	10	0	5	10
Recognition Accuracy (%)	95	89.83	89.83	95	89.83	89.83	33.16	33.16	32.5	32.67	33.16	32.5	57.14	79.20	90.40
Word Error Rate (%)	5	10.17	10.17	5	10.17	10.17	66.84	66.84	67.5	67.33	66.84	67.5	42.86	20.8	9.6

The performance comparison results as given in table 5.4 can be best understood in figure 5.11.

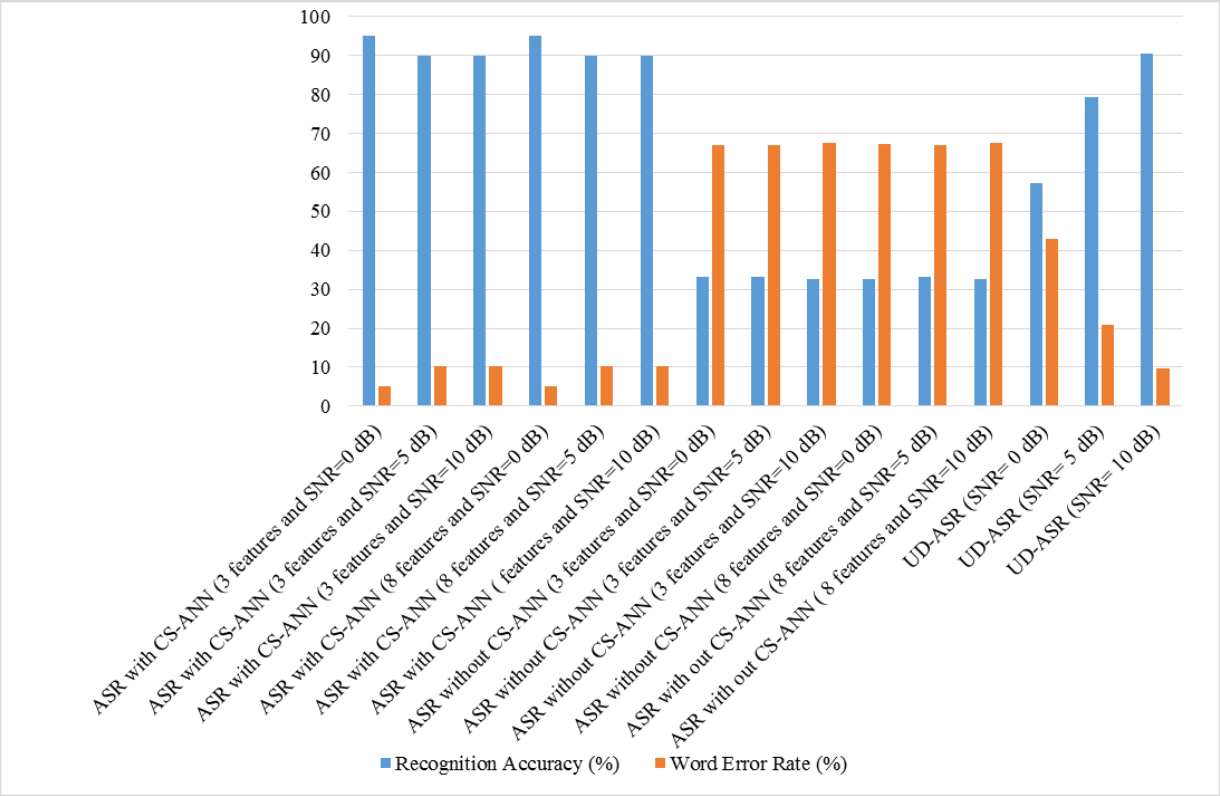


Figure 5.11: Performance Comparison of Proposed Methodology (ASR with CS-ANN), ASR without CS-ANN and Uncertainty decoding based ASR

From the results of the proposed methodology as well as its comparison with the existing techniques the recognition rate achieved by this system is better than other methods such that the recognition accuracy is 95% with zero levels of Babble noise and even after increasing the noise levels to 5 and 10 dB the accuracy of recognition is better than the method without optimization. Similarly, the UD-ASR method has the recognition accuracy of 57.14%, 79.20% and 90.40% with noise levels of 0 dB, 5 dB and 10 dB respectively and the discussion about these results is given in the following section.

5.3.5 Result Analysis

The results are shown in tables 5.2-5.3 and figures 5.7-5.11 depict that better performance results are obtained with the proposed man-machine interaction system for ASR compared with other techniques. The recognition rate achieved with this system is 95% which is better than the existing method and without CS-ANN. This is depicted in table 5.3 as well as in

figure 5.8, such that the system without CS algorithm in optimizing the ANN yields poor recognition results than the system with CS as well as the existing method. Though the accuracy of the proposed method is superior as compared to the method which is taken into consideration even in the presence of babble noise at different levels, the method achieved lower results compared to this proposed here in lower SNR levels. The recognition results also show that most of the words in the testing phase are identified clearly and hence the lower word error rate. Hence the proposed Fuzzy based DWT feature extraction produces better recognition and accuracy of the ASR system at low SNR levels and the accuracy level is maintained even the SNR level is increased which demonstrates the efficiency of the suggested method in achieving better accuracy results. The comparison result is given in table 5.3. It also presents that with the use of the suggested feature selection approaches the accuracy levels remain stable and with the CS-ANN classification the accuracy level is increased. In addition to that, the employment of the optimization algorithm and the extracted features also have a major impact on recognizing the words as given by the comparison results.

5.4 CHAPTER SUMMARY

In this chapter, a method for the man-machine interaction system in ASR systems through fuzzy-based feature selection and ANN optimized with CS optimization algorithm.is suggested. The speech signals are initially converted into samples, and frames using the Hamming window and the noise levels are suppressed by harmonic decomposition. The next eight different features are extracted from the speech signal through DWT and the most relevant features are selected using fuzzy logic. The selected features are employed in training the NN in which the optimization of the network is performed by the CSO algorithm. The experimental results are presented and compared with conventional technologies under various conditions and the results show the efficiency of the proposed methodology. The experimental results given in section 4.3 reveal that the proposed ASR can achieve the recognition accuracy of 95% with a lower word error rate of 5% which shows the betterment of the proposed work. The recognition accuracy achieved by this method is 95% with SNR level of babble noise at 0 dB, 89.83% at both the SNR levels of 5 dB and 10 dB respectively and it is a better result compared to the uncertainty decoding based method.

CHAPTER 6.

CUCKOO SEARCH OPTIMIZATION BASED ANN CLASSIFIER

6.1 INTRODUCTION

Automatic speech recognition (ASR) innovation has progressed quickly in the past decade. While numerous ASR applications utilize effective computers to handle complex recognition algorithms, there is an interest in a successful solution on installed frameworks like communication equipment and several low-cost consumer electronic systems. Speech recognition has made boundless innovative advances in numerous fields, for example, call directing, programmed interpretations, data seeking, data entry and so on. Speech recognition has been achieved by consolidating different algorithms drawn from diverse disciplines, for example, statistical pattern recognition, signal processing and phonetics and so on[30]. Even after years of research and numerous effectively deployed commercial products, the execution of programmed speech recognition frameworks in genuine utilization situations lags behind human-level execution. Speech recognition is a pattern classification issue and speech recognition frameworks utilize detached word recognition.

Speech is a composite signal that principally conveys data about the message to be passed on, speaker attributes, and dialect. Speech signals contain a huge measure of data and can be depicted as having different levels of data. At the top level, there are lexical and syntactic features, underneath that are prosodic features, further beneath these are phonetic features, and at the most essential level, low-level acoustic features are present, which by and large give data on the framework that makes the sound, for example, the speakers' vocal tract [137]. The qualities of a speech signal can be credited to the measurements of the vocal tract framework, attributes of excitation, and the learning propensities of the speakers. The speech data is represented by spectral features like mel-frequency cepstral coefficients (MFCCs) and linear prediction (LP) cepstral coefficients. Efforts are being made to exploit the effectiveness of features extracted from excitation source attributes and suprasegmental qualities for speaker recognition.

A lot of research has been done in the field of pattern recognition which has resulted in the development of some good classifiers like the Hidden Markov Model, Gaussian Mixture Model, Neural Networks, Support Vector Machine and so on. Artificial Neural Networks

(ANN) are biologically stimulated tools for data processing [138]. The multilayer feed-forward networks, the recurrent network and so on can be trained to associate input data, to learn anonymous words. Speech recognition displayed by artificial neural networks (ANN) doesn't oblige from the earlier learning of speech procedures and this system rapidly turned into an attractive alternative option for HMM. Three main factors were accountable for the recent development of neural networks as high-quality acoustic models they are creating the networks deeper makes them more powerful, therefore deep neural networks (DNN), initializing the weights reasonably and utilizing much faster hardware creates it possible to train neural networks efficiently, and utilizing a larger number of context-dependent output units significantly increases their performance.

6.2 PROPOSED ASR SYSTEM

Speech recognition is the process of conversion of spoken words into text by the computer system. In this research work, a novel automatic speech recognition (ASR) system to provide interaction between humans and machines is developed. The proposed ASR system consists of three important steps. They are signal preprocessing, feature extraction and classification. The process flow of the proposed speech recognition is shown in figure 6.1.

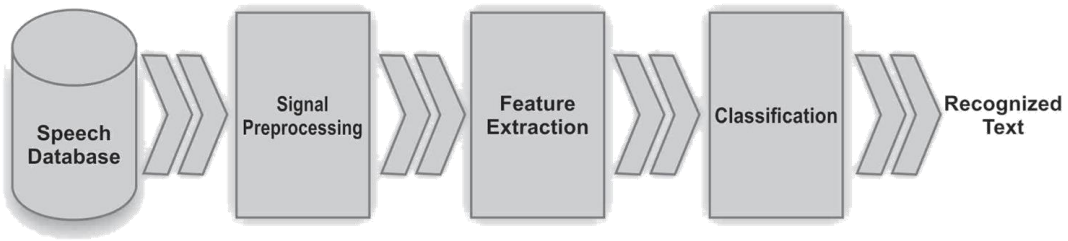


Figure 6.1: Process flow of the Speech Recognition System

Preprocessing is done to remove the noise present in it. In feature extraction, two kinds of features are extracted from the pre-processed speech signal. Then these extracted features are given to the classification phase here Artificial Neural Network is used as a classifier. To improve the execution time as well as to improve the recognition performance here the weights in the neural networks are optimized by using an evolutionary algorithm called Cuckoo Search Algorithm. Then the recognized text related to the corresponding input speech signal is obtained as output.

6.2.1 Speech Signal Preprocessing

Generally, the input speech signals have noise, which is created through fans, computers, the conversation between people and so on. These kinds of interruptions will affect the performance of the speech recognition process. Therefore, the first important step in speech recognition is the preprocessing of the speech signals which is executed to remove avoidable waveform of signal and to truncate the task of recognition. In this preprocessing, the Speech signal from the database is passed to the Moving average, High pass filter for removal of background noise. After removing the background noise, the speech signals are framed & passed through the window. The output of the window has a pre-processed signal & can be used further for feature extraction.

6.2.2 Feature Extraction

Feature extraction is the process of extracting the most important features from the signal to recognize them. In this research work, two kinds of acoustic features are extracted from the speech signal they are Mel Frequency Cepstrum Coefficients (MFCC) and Linear Predictive Coding Coefficients (LPCC).

(i) Mel Frequency Cepstrum Coefficients (MFCC): Calculation of MFCC of the speech signal is based on the short-term analysis, and hence from each frame, the MFCC feature vector is calculated. To extract the coefficients, the pre-processed speech sample is taken as the input and is separated into several frames. Afterwards, the hamming window is applied to decrease the discontinuities between the frames. Then, the Discrete Fourier Transform (DFT) is utilized to generate the Mel filter bank. As indicated by Mel frequency wrapping, the width of the triangular filters differs thus the log total strength in a discriminating band around the centre frequency is incorporated. After warping, the quantities of coefficients are acquired. At last, the Inverse Discrete Fourier Transform (IDFT) is utilized for the cepstral coefficient computation. It changes the log of the quefrench space coefficients to the frequency domain.

Finally, a total of 13 coefficients are extracted from the given sample and it is used further for the classification process.

(ii) Linear Predictive Coding Coefficients (LPCC): The LPCC analysis of a speech sample can be estimated as a linear grouping of previous speech samples. LPCC is a frame-based analysis of the speech signal which is executed to provide observation vectors of the speech

sample. To calculate LPCC features, first, the speech signal is divided into frames of samples. Every frame is multiplied by a sample Hamming window, and this windowed frame is delivered to perform short term autocorrelation. Then, LPCC analysis is implemented based on the Levinson-Durbin recursion. It gives a 2Qby-T matrix of observed features. The LPCC coefficients are then transformed into Q cepstral coefficients, which are biased by a raised sine window. The first half of an observed vector is the weighted cepstral sequence for a particular frame, the second part is the time difference weighted cepstral coefficients which are utilized to add dynamic information. Here, 24 feature vectors are extracted using LPC analysis.

Now, the two extracted features are given to the classification phase consisting of the Cuckoo Search Optimization-based Artificial Neural Network to recognize the corresponding text.

6.2.3 Classification by CSO Based ANN

In this proposed method, the artificial neural network is used as a classifier.

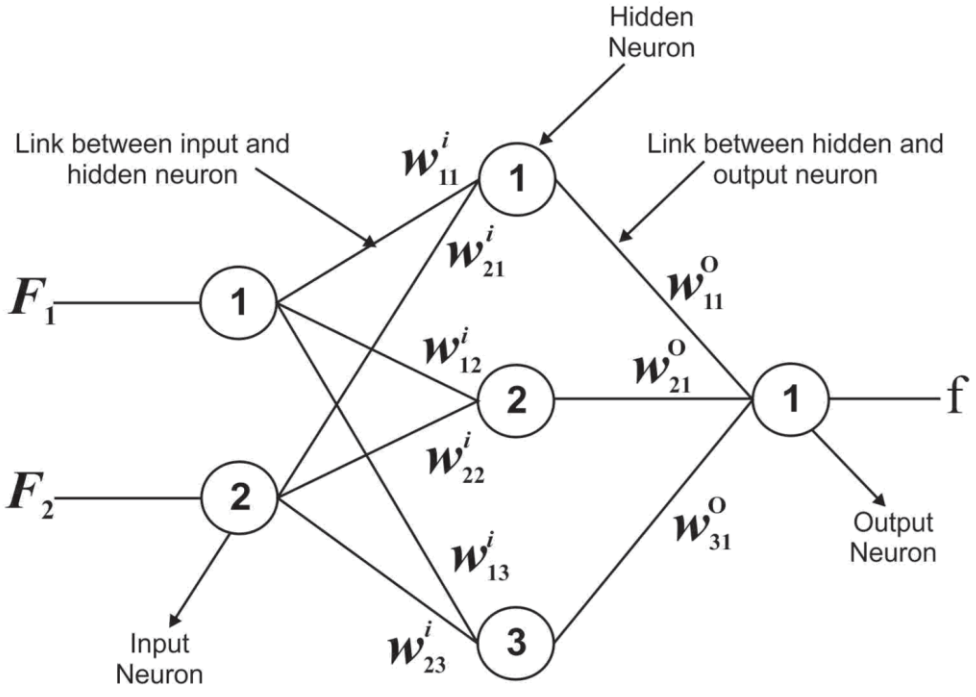


Figure 6.2: Structure of Neural Network with two input features

The neural network is a three-layered classifier with 'n' input nodes, '*l*' hidden nodes and 'k' output nodes. In this proposed work, a two-layered Feed Forward Backpropagation

Neural Network (FFBNN) with 3 units; two input units, three Hidden units, and one output unit is implemented. Figure 6.2 gives the structure of the FFBNN classifier. Here, the input layer consists of two inputs having two features extracted which are MFCC and LPCC features. These features are given as input in which networks get trained and produce corresponding output. In neural network weights between the layers are assigned by using the Cuckoo Search Optimization (CSO) algorithm. These weights are randomly assigned and CSO assigns the best weights to correctly classify the given input.

The steps to implement an artificial neural network are shown in figure 6.3.

1. Assign the weights for neuron.
2. Generate the neural network with extracted features (F_1, F_2) as input units, H hidden and f as output unit.
3. Calculation of the bias function for input layer is given by

$$X = \sum_{n=1}^{3} w_{1(n)} F_1 + w_{2(n)} F_2$$

4. Calculation of activation function for output layer is given by

$$Active(X) = \frac{1}{1+e^{-X}}$$

5. Identify the learning error as given below

$$Learning\,Error = \frac{1}{NF} \sum_{n=0}^{NF-1} Y_n - Z_n$$

Where Y_n - Desired output, Z_n - Actual Output and NF – Number of features.

Figure 6.3: Steps in Neural Network

In Neural Network, Back Propagation Algorithm is utilized as the Learning Algorithm. To create the training set, it wants a dataset of the required output for different inputs. Generally, the Back Propagation Algorithm is useful for Feed-Forward Networks. This learning algorithm needs that the activation function utilized by the neurons be differentiable. Simultaneously the weights for the neurons of the hidden layer and the output layer were assigned using the Cuckoo Search Optimization algorithm.

CSO is a population-based optimization algorithm, and like other meta-heuristic algorithms, it begins with an arbitrary initial population. CSO algorithm essentially works

in three stages: choice of the best source by keeping the best nests or solutions, supplanting of host eggs concerning the quality of the new solutions or cuckoo eggs created based randomization utilizing Levy flights all around and discovering cuckoo eggs by the host flying creatures and supplanting as per the quality of the neighborhood random walks. Every cycle of the search comprises a few stages of initialization of the quality nest or solution, the quantity of accessible host nests is settled, and the egg laid by a cuckoo is found by the host feathered creature with a probability $p_\alpha \in [0,1]$.

In the proposed method, each best nest signifies a possible solution that is the weight space and the equivalent biases for neural network optimization. The weight optimization issue and the size of the population denote the quality of the solution. In the first epoch, the finest weights and biases are initialized with CSO, and these weights are given to the neural network. The weights in the neural network are computed and compared with the finest solution in the backward direction. In the next cycle, CSO will update the weights with the finest possible solution and CSO will continue searching for the finest weights until the last cycle/ epoch of the network is reached.

The pseudocode of the proposed algorithm is shown in figure 6.4

Step:1 CSO Initialized and pass the finest weights to neural network.
Step:2 Training data loaded
Step:3 **While** (Mean Squared Error) MSE<stopping criteria
Step:4 Initialize the cuckoo nests
Step:5 Cuckoo nests are given as weights to network
Step:6 Neural network runs utilizing the weights initialized with CSO
Step:7 Now calculate the error backward
Step:8 CSO keeps on computing the finest possible weight at each iteration until it meets network convergence
End While

Figure 6.4: Pseudocode for Proposed Algorithm

6.3 EXPERIMENTAL RESULTS

The proposed system is implemented in the working platform of MATLAB with the following system specification.

Processor : Intel i5 @ 3GHz

RAM : 8GB

Operating system : Windows 8

MATLAB version : R2013a

The speech database consists of 465 different words collected from different individuals from India with an age group of 19 to 40 years. The dataset was recorded by using Audacity 1.3 beta. The training data consists of 12,987 utterances spoken by 20 males and 17 females. Here, a sample signal is taken from the database to validate the proposed method and the speech signal is shown in figure 6.5.

Then by applying the preprocessing technique, the noise in the signal is removed and the pre-processed signal is shown in figure 6.6.

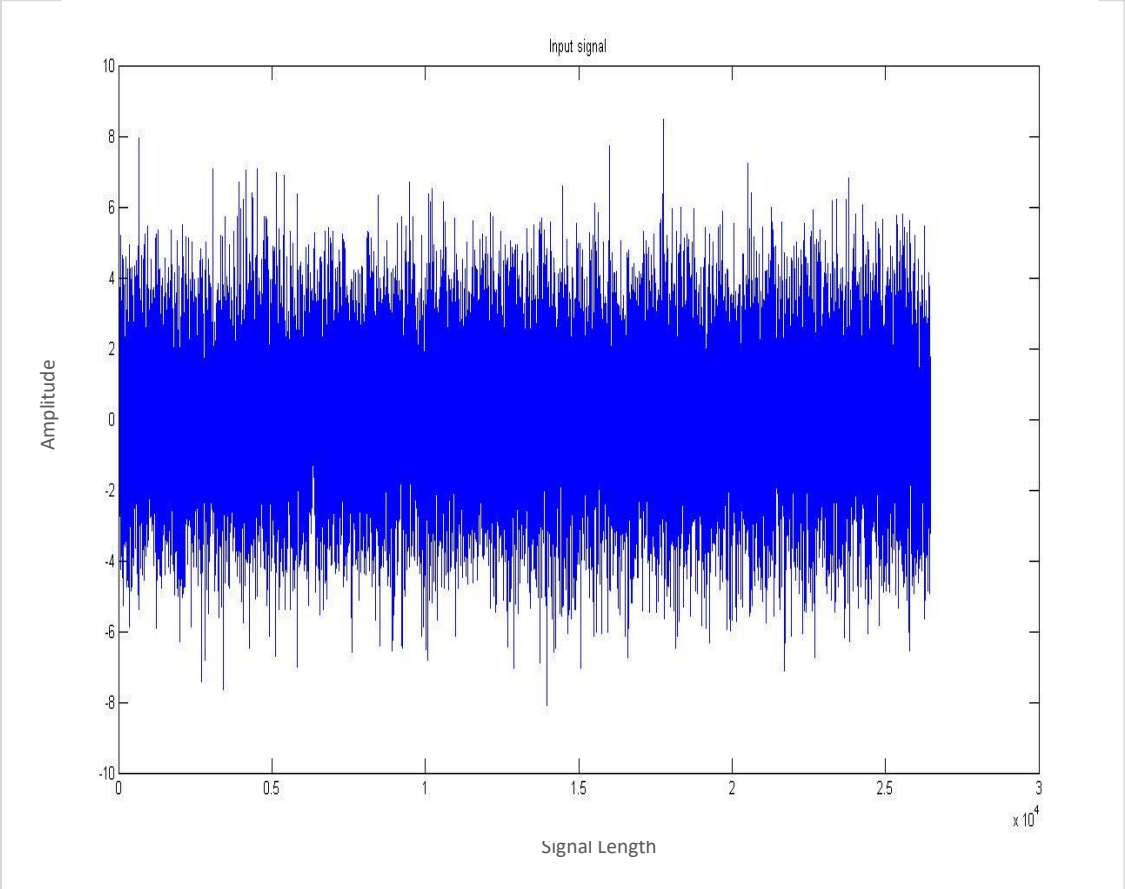
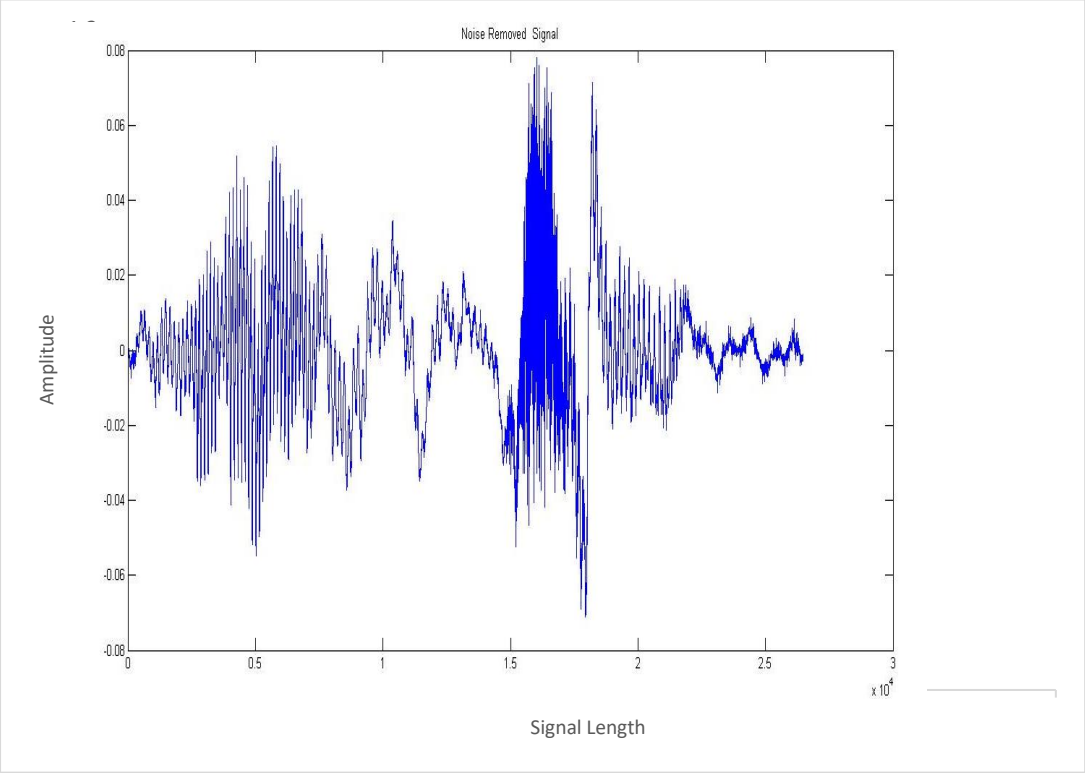


Figure 6.5: Input Speech Signal

The figure above shows one input signal from the database. The results are validated using this input speech signal. Then by applying the preprocessing techniques and noise removal etc., the pre-processed signal is obtained as shown in figure 6.6. This

signal is used for feature extraction. The extracted features are given to the neural network classifier to give the desired text output. Thus, the conversion of a speech



signal to the text output has been carried out.

Figure 6.6: Pre-processed Speech Signal

Figure 6.6 shows the pre-processed signal as extracted from the input signal. From the pre-processed signal the MFCC and LPCC features are extracted, and their respective values are shown in table 6.1.

Table 6.1: Feature Values

S. No.	Features	Value
1	MFCC	0.18665
2	LPCC	-0.23001

These features are given to the neural network-based classifier. Based on the features extracted and the classifier used, finally, the recognized text is produced at the output of

the system. The two features extracted are the MFCC and the LPCC. The extracted features are sent to the classifier for further recognition purposes

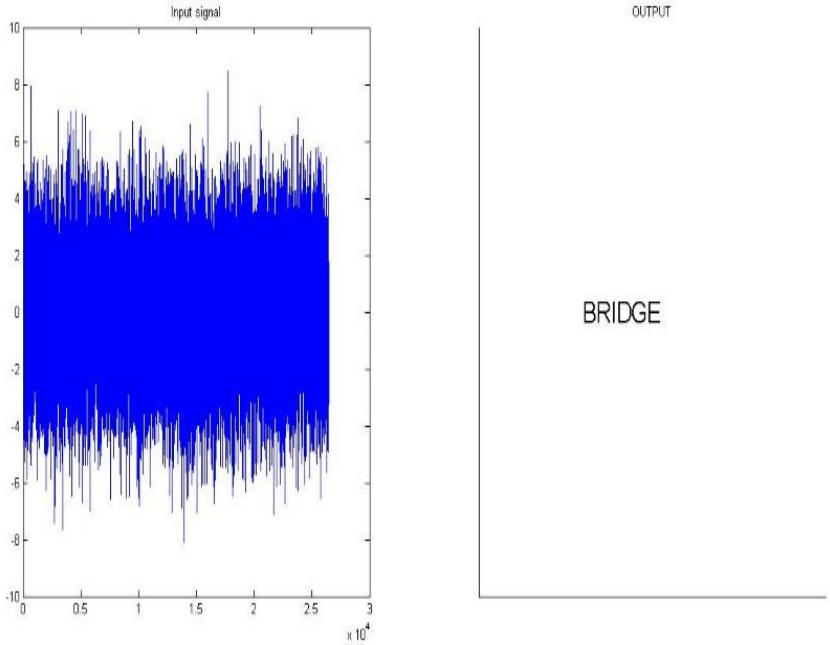


Figure 6.7: Output of the ASR System

The convergence graph of Cuckoo search optimization is shown in figure 6.8.

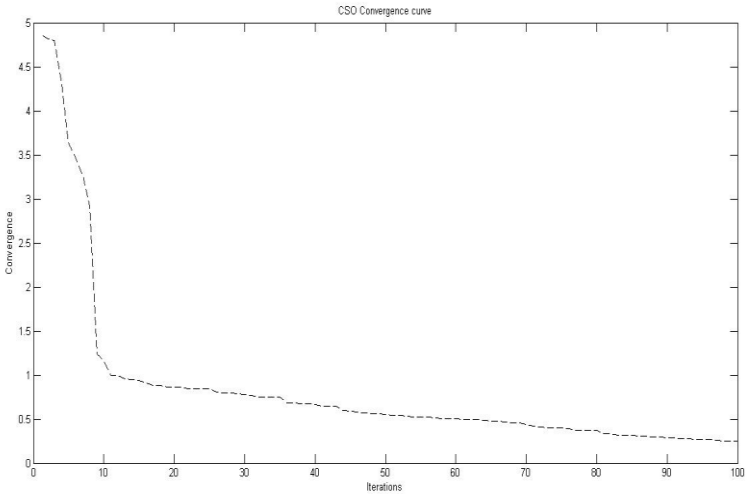


Figure 6.8: Convergence graph of Cuckoo search optimization

The recognition accuracy performance (with CSO based ANN) is compared, and the results are listed in table 6.2 below. From table 6.2, it is seen that the recognition accuracy of the

proposed CSO based ANN has better accuracy compared with the other two existing methods. From this, it can be concluded, that the weight optimization by CSO in a neural network has increased the performance of the system.

Table 6.2: Recognition Accuracy

S. No.	Method	Accuracy
1	Proposed CSO-ANN	89.65%
2	GA-ANN	82.78%
3	ANN	74.30%

6.4 CHAPTER SUMMARY

The backpropagation algorithm is one of the widely used techniques to optimize the feed-forward neural network training. The conventional algorithm had some disadvantages, such as being stuck in local minima and low speed of convergence. In this chapter, a novel meta-heuristic algorithm, called cuckoo search optimization (CSO) is proposed to train the neural network to achieve a fast convergence rate and to increase the recognition accuracy of the speech recognition system. Here, two kinds of features are extracted Mel Frequency Cepstrum Coefficients (MFCC) and Linear Predictive Coding Coefficients (LPCC) which are given to artificial neural networks for the training process. The experimental results show that the proposed CSO based ANN is far better than the existing methods in terms of convergence rate, simplicity, and accuracy.

CHAPTER 7.

ASR MODEL FOR NEWS TRANSCRIPTS

7.1 INTRODUCTION

Automatic speech recognition is studied for the benefit of application users. A distant microphone system is an instance that is widespread among household users. In today's living system, a high level of background noise is encountered that develops numerous challenges for recognition systems [139]. Though the disturbing noise is recognized, in some scenarios, the performance of intelligent recognition systems is not enhanced. Automatically, it interrupts the speech signal, and this noise-masking causes phonetic information loss. Thus, the recognition error rate is high due to the misinterpretation of the speech signal. In a few cases, misinterpretation due to the sensitivity of the speakers degrades the effects of speech production and noise masking. This is being resolved by the automatic speech recognition system that aimed for proper handling of speech and its background noise [140]. The system generally converts the speech signal into text. Since it's a developing field, the research on man-machine interaction will provide novel insights. Speech signals are one of the complex signals which make the Automatic Speech Recognition Systems (ASRS) field, an innovative process.

The robust feature extraction process [142] determines the success rate of intelligent speech recognition systems. The ideal representation of the feature vector should carry an error-free speech signal. Hence, the examination of robust feature extraction has been major research in the ASR field. Though, there are several robust feature extraction techniques are available from past decades, the common issues like bandwidth and the number of filters, matter the most. Generally, types of ASR systems are categorized based on the number of speakers, utterance nature, vocabulary size, and bandwidth. Understanding the production of speech sounds and the source of variability determines the level of ASRS functionalities [143]. Unlike text, phonetics (or) words do not have boundaries due to the incessant flow of speech. This simple sample defines the difficulty level of speech interpretation, an utterance of six sheep may be sick sheep, even in the absence of noise. Also, in a printed text, multiple occurrences of a letter appear the same. In contrast, spectral and temporal characteristics of speech sound vary a lot, which depends on several factors.

ASR is useful when speech articulation errors are consistent and predictable. It is extremely useful when the speech sounds might seem unintelligible to human ears. There are not many products available to process text which has been uttered unrestrictedly [144]. Moreover, a larger amount of training data is required to train the ASR systems that are meant for people with speech difficulties like Dysarthria. These people are linguistically correct, but the articulation of speech is unclear. This chapter is divided into the subheadings and the related work in this field, the proposed methodology, experimental analysis, and conclusion.

7.2 PROPOSED METHODOLOGY

This section presents the proposed model of the research study. The survey states that the analysis is done in Automatic Speech Recognition (ASR) is still a challenging and demanding task. Due to variation occurring in word utterance, this recognition system is in the upcoming research area. Feature extraction plays a vital role in ASR systems because speech quality purely depends on it. Though variant models were suggested by the scholars, the accurate translation of the speech text from the speech signal is not yet achieved. Most of the analysis is done using Hidden Markov Models which throws issues like recognition error rate, accuracy, and speech perplexity. Here, an attempt to resolve the above-mentioned issues and achieve better translation and recognition rates than others is done. The proposed phases are explained as follows:

7.2.1 Data Collection

This is the first step of the study. The news report dataset comprises audio and text records that were collected from a public repository. Since it’s a news dataset, the acquisition of relevant and irrelevant data is higher i.e., missing fields, irrelevant data entered irrelevant attributes.

7.2.2 Pre-processing

Pre-processing is one of the significant steps in this study. Here, a simple normalization technique is applied to both audio and text records. It is a process that alters the range of intensity values. It is also known as contrast stretching. The consistency of the signals is maintained by maximum and minimum values under dynamic ranges. According to equation 7.1, the normalized values are computed as follows:

$$I_n = (I - Min)\,[(newMax - newMin) / Max - Min] + newMin \quad (7.1)$$

7.2.3 Feature selection

This is the most important and core part of the automatic speech recognition systems because it depicts the quality of the speech signal. Mel Frequency Cepstral Coefficient (MFCC) is employed for extracting the features from signals. Figure. 7.1 presents the working flow of MFCCs.

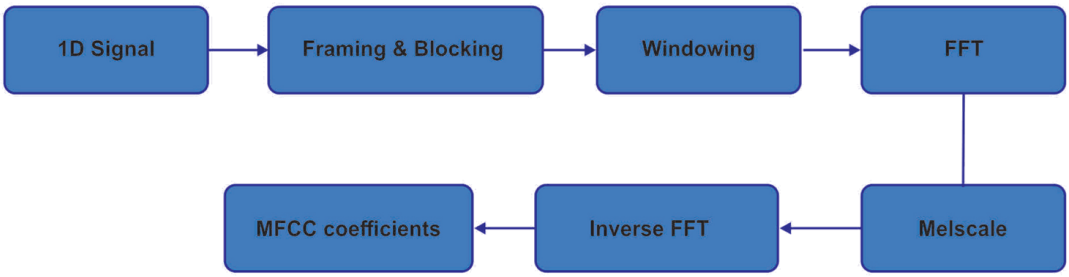


Figure 7.1: The working flow of MFCCs

7.2.3.1 *Framing and blocking:*

The collected 1D signal is segmented into n frames of N samples. Likewise, the M samples in adjacent frames overlapped by N-M samples. If the frame size is smaller than the sample sizes, then enough details are not acquired from the frame. Until the signal ends, the frames are processed.

7.2.3.2 *Windowing:*

In this step, the colliding of frames is minimized. Mostly, hamming windows are used for representing the input and output signal. The eqn.3.2 presents the estimation of frames in windows.

$$W_n\,(m) \;=\; 0.54 \;-0.46\,cos\,(2\pi m\,/\,N_m - 1) \quad (7.2)$$

7.2.3.3 *Fast Fourier Transform (FFT):*

It converts the spatial domain into the frequency domain. Each frame holds samples that convert into a frequency domain. It is one of the fastest algorithms which applies Discrete Fourier Transform (DFT).

$$D_k \;= \sum_{m=0}^{N-1} D_m\, e^{\frac{-j2km}{Nm}} \quad (7.3)$$

Where, k= 0, 1, 2.... N_{m-1}

Each DFT is computed separately for an easier computation process. In digital processing, the other area directly uses DFT.

7.2.3.4 *Mel Scale:*

Here, a Triangular filter bank is utilized for computing energy function. It is a set of bandpass filters that are being decided by steady Mel frequency time. When frequency gets higher, the Mel space becomes wider.

Mel-scaling mapping is done between observed frequency scales (Hz) and the perceived frequency scale (Mels). Finally, the first 13 coefficients were used, and the higher values are eliminated. These computed coefficients are then fed into DCT.

7.2.3.5 *Discrete Cosine Transform (DCT):*

It helps to convert the log Mel spectrum into the spatial domain. The DFT is desirable for all coefficients estimation from DCT. The output obtained after applying DCT is known as Mel Frequency Cepstral Coefficient.

$$C_n = \sum_{m=0}^{k-1} (\log D_k) \cos [m\ (k-1)\ \pi] \quad (7.4)$$

Where *m*=0, 1...k-1

C_n= MFCC value

m= No. Of coefficients

7.2.3.6 *Mel Frequency Cepstral Coefficient:*

The MFCC converts the input signal from 2D to 1D signal. This 1D signal is being categorized into frames and the neighbouring frames are represented as ($M<N_m$). By doing so, the loss of information is avoided. At windowing, the frequency domain is achieved by FFT. Then, the estimation of the magnitude spectrum is further transformed by Mel frequency. For audio signals, the features retrieved are coefficients, delta, and the delta-delta. Likewise, the text features are mean, entropy, variance, skewness, kurtosis, and the root mean square.

7.2.4 Classification

It is one of the core parts which takes features, as input, to CNN and then speech is recognized in text form. It comprises three layers, namely, the input layer, an output layer, and several hidden layers. Each layer has its functioning process with learning features. The most common layers in CNN are convolution, activation or ReLU, and pooling.

(i) Convolution: A set of convolutional filters is used for activating certain features.

(ii) Rectified Linear Unit (ReLU): It maps the negative values to zero and administers the positive values. These are further processed onto activation layers.

(iii) Pooling layer: It identifies the variant features by nonlinear downsampling.

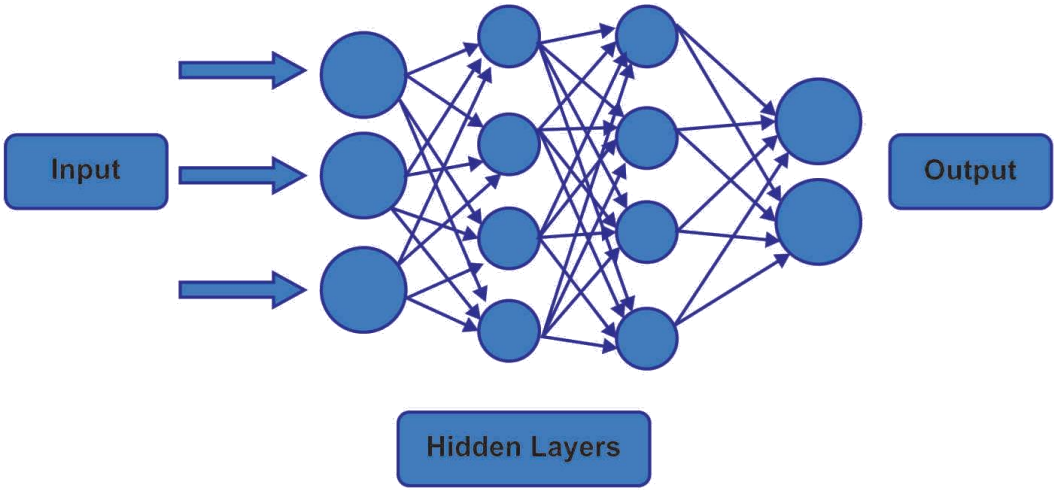


Figure 7.2: Working of CNN

Once receiving the learning features, the CNN helps for classification. Since it is a connected layer, the k dimensions define the number of classes.

Figure 7.2 describes the working of the Convolution Neural Network.

The special structure such as local connectivity, weight sharing, and pooling in the Convolution Neural Networks exhibits some degree of invariance to small shifts of speech features along the frequency axis, which is important to deal with speaker and environment variations.

The input contains several localized features organized as some feature maps. The size (resolution) of feature maps gets smaller at the upper layers as more convolution and pooling operations are applied.

The CNN has three key properties: locality, weight sharing, and pooling. Each one of them has the potential to improve speech recognition performance.

Figure. 7.3 represents the workflow of this study.

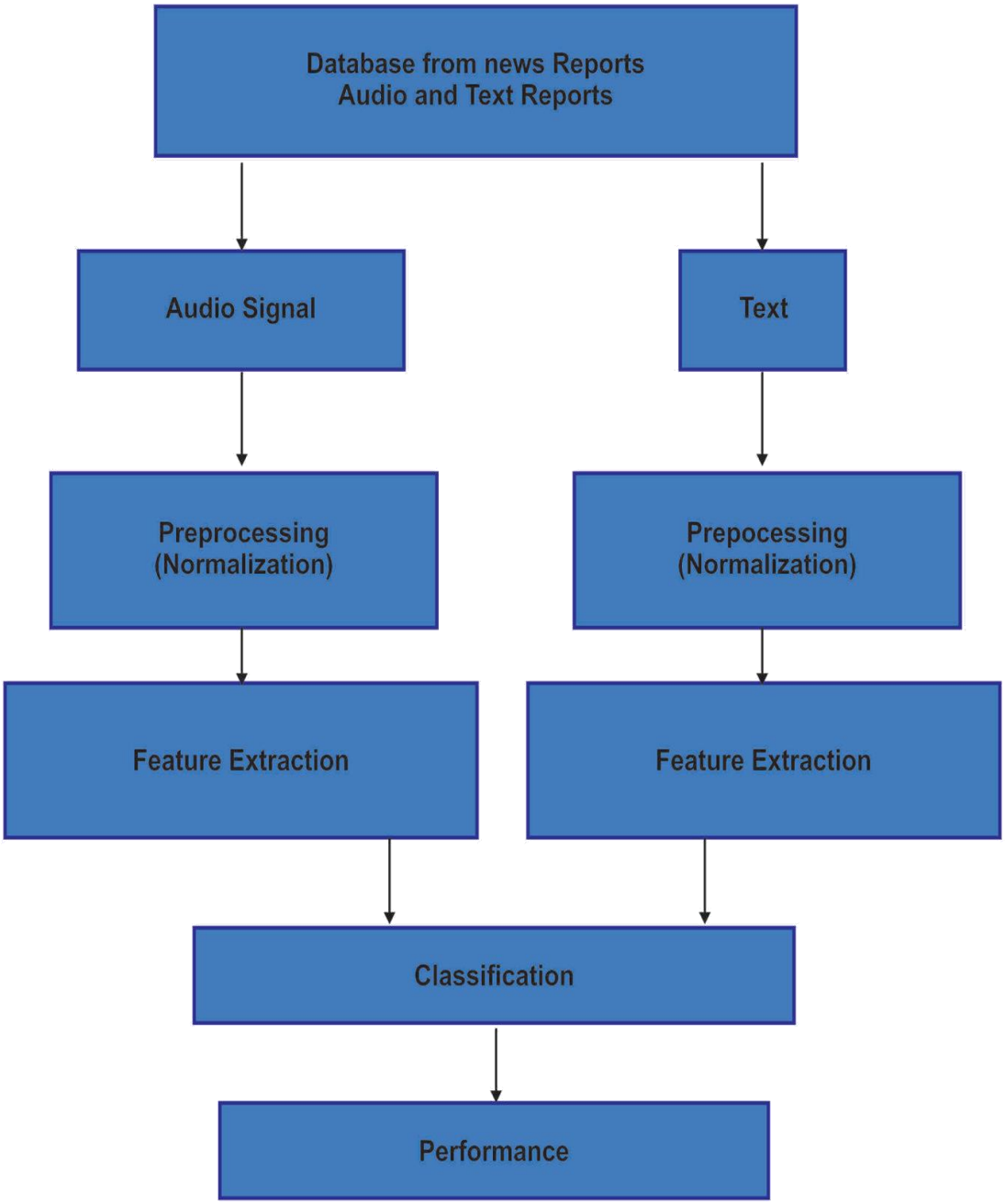


Figure 7.3: Proposed Workflow

7.3 EXPERIMENTAL RESULTS

This section presents the experimental analysis of the proposed model. The proposed CNN classifier is experimented with using MATLAB, a programming language.

Initially, the data is being collected from a public repository. The confusion matrix is generated for recognizing its relevant classes. The performance of the CPU is taken while training the features.

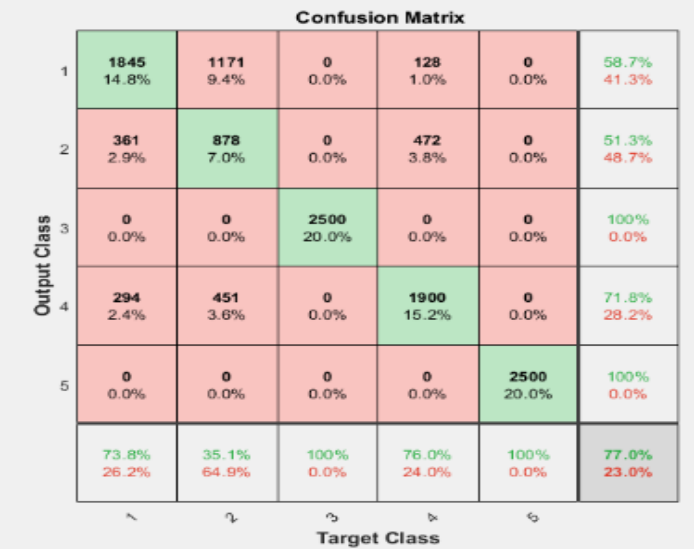
CNN's have been applied to speech recognition in a novel way, such that the CNN's structure directly accommodates some types of speech variability. A performance improvement convolving along the frequency axis creates a degree of invariance to small frequency shifts, which normally occur in actual speech signals due to speaker differences.

Table 7.1: Performance of CPU for training the features.

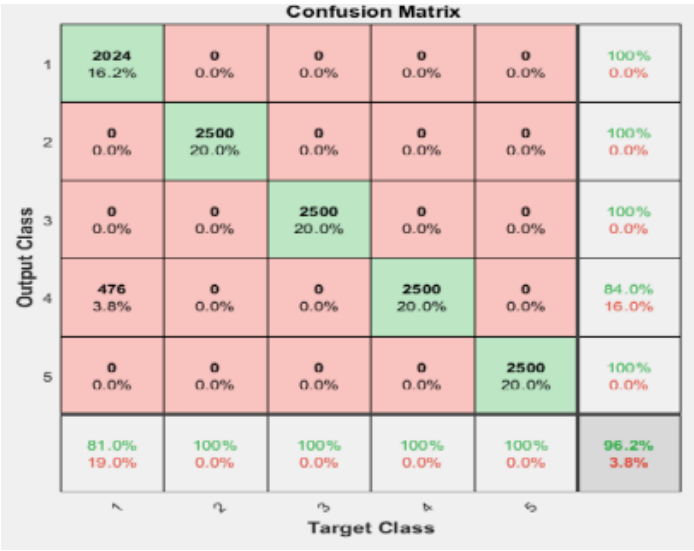
Epochs	Iteration	Elapsed time (s)	Accuracy (%)	Loss	Learning rate
1	1	0	0	1.6094	0.0100
1	50	0	100	0.2267	0. 0100
1	100	1	100	0.3115	0.0100
1	150	1	100	0. 1779	0.0100

The training of features at iterations of 1, 50, 100, and 150 are performed and the elapsed time is noted. The accuracy and the minimized loss rate are determined and are also shown in the table. The network learning rate is also shown.

The results depict that the minimized loss rate helps in achieving better accuracy. It is very evident from table 7.1.



(a): Confusion matrix of existing ANN

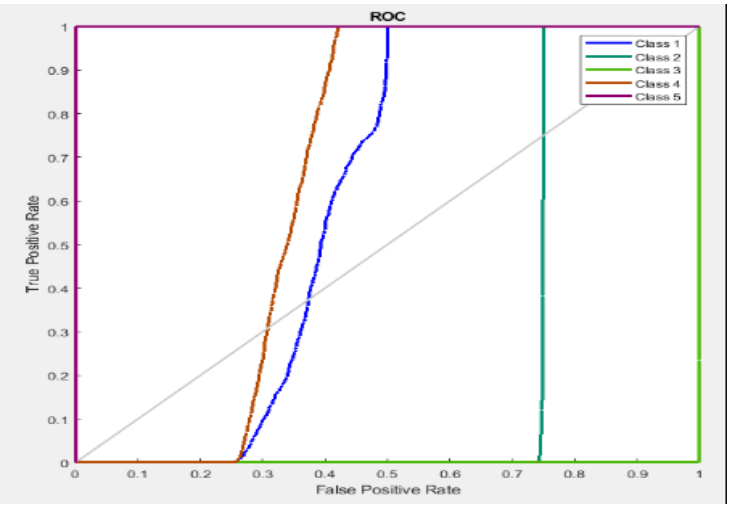


(b): Confusion matrix of proposed CNN

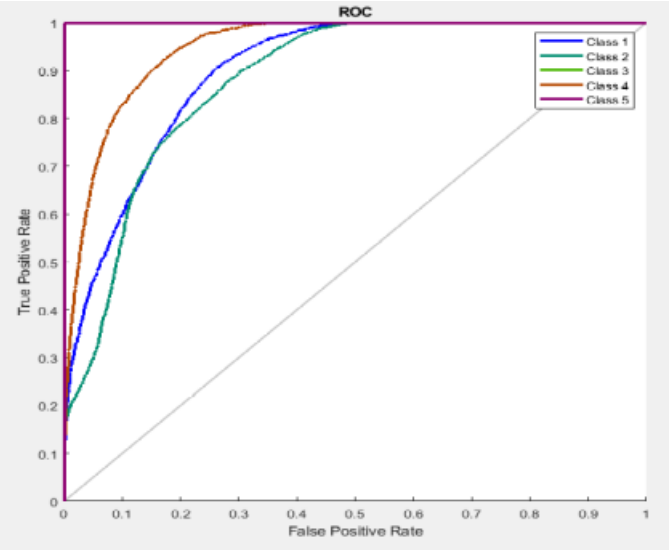
Figure 7.4: Confusion matrix of (a)existing ANN (b)proposed CNN

Figure 7.4 presents the confusion matrix between the existing ANN and the proposed CNN. The matrix analysis depicts that the loss incurred by CNN is better than ANN. And thus, the minimized loss rate helps in achieving better accuracy. Thus, the proposed CNN is a better classifier than ANN.

The Region of Convergence Curve (ROC) defines the area covered by the bounded region. The below figure.7.5 compares the ROC between existing ANN and proposed CNN. Figure 7.5(a) represents the ROC of the existing ANN and figure 7.5(b) represents the ROC of the proposed CNN.



(a): ROC of Existing ANN



(b): ROC of Proposed CNN

Figure 7.5: ROC (a) Existing ANN (b) Proposed CNN

Figure .7.5 presents the comparison of ROC between the proposed CNN and the existing ANN. It depicts that the proposed CNN intelligently classifies its class better than the existing ANN. The performance metrics studied are explained as follows:

i) F-measure: It conveys the balance between precision and recall which is given in the equation. 7.7.

$$\text{F-measure: } 2^* ((\text{precision} * \text{recall})/ (\text{precision} + \text{recall}) \quad (7.7)$$

ii) Accuracy: It defines the ability to distinguish between normal and abnormal cases. It is given in equation 7.8.

In speech recognition, it would be the distinction between correctly recognized words and the errors or wrong output words.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (7.8)$$

iii) Sensitivity: It defines the ability to distinguish abnormal samples correctly. It is given in equation 7.9. It defines how sensitive the system remains. That is, if it can distinctly identify the wrong words.

$$\text{Sensitivity} = (TP) / (TP + FN) \quad (7.9)$$

iv) Specificity: It defines the ability to estimate normal samples correctly. It is given in equation 7.10.

$$\text{Specificity: } (TN) / (TN + FP) \quad (7.10)$$

7.4 CHAPTER SUMMARY

The recent innovation in information processing systems combined with intelligent models has gained much interest among researchers. Since speech is a complex process that demands the coordination of articulation, breathing, and facial expression. This research attempts to intelligently recognize the text from speech signals using improvised convolutional neural networks. The required data is collected from a public repository, a news report dataset that generally comprises irrelevant data. It is being pre-processed by min-max normalization techniques that efficiently normalize the intensity value of the data. The normalized data is then applied using MFCC which depicts the relevant features of audio and text data. The features extracted from audio are coefficients, delta, and delta-delta, and the text features are mean, entropy, variance, skewness, kurtosis, and the root mean square. These extracted features are fed into Convolutional Neural Networks (CNN) and the relevant signals are classified. The proposed model is investigated over the News Report dataset that composes of two sources of data, audio, and text. The proposed CNN is compared with the existing ANN which proves the efficiency in terms of accuracy, sensitivity, specificity, precision, recall, f-measure and the gaussian mean (gmean). The accuracy of the proposed CNN 96% achieved is better than ANN, 76%.

Table 7.2: Performance Comparison between existing ANN and proposed CNN.

Classification	Accuracy	Sensitivity	Specificity
CNN	96.170	80.84	100
ANN	76.98	73.80	77.78

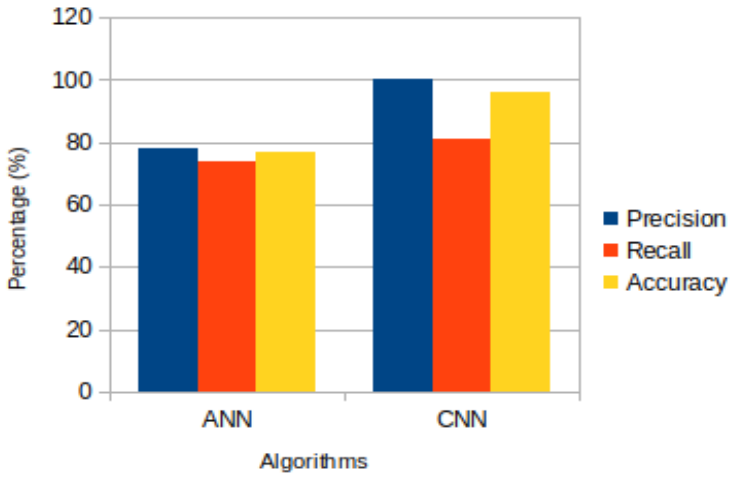


Figure 7.6: Performance Graph between existing ANN and proposed CNN.

CHAPTER 8.

CONCLUSION AND FUTURE SCOPE

8.1 CONCLUSION

The current work is mainly concentrated on both acoustical and perceptual analysis of intonation. To get measurable data, it has been proposed to follow the scientific method. It is easy to implement this acoustically measurable data into computer algorithms especially for Text to speech Systems and Automatic speech recognition systems. The knowledge can ideally be used in speaker identification because intonation is the intrinsic property of a speaker, and it will be able to distinguish the speech of one speaker from others.

8.2 PERFORMANCE COMPARISON OF DEVELOPED ASR SYSTEMS

The researchers in [143] and [146] have suggested an isolated ASR and a fuzzy-based ASR via two distinct techniques for the feature extraction procedure. One of the techniques uses ANN with Back Propagation (BP), whereas the other used a fuzzy-based Discrete Wavelet Transform (DWT) for feature extraction. In the ANN-based technique, the audio signal is pre-processed to minimize the noise and after this, the general features like word size, sampling point, pitch and some statistical parameters like kurtosis, variance, skewness, entropy, and mean are extracted. After extraction, these features are augmented for training the classifier. In the end, the feed-in audio signal is exhibited as text output. On the other hand, in the fuzzy based technique, the feed-in audio signal goes through pre-processing which includes sampling of the signal to generate frames by Hamming window technique. By employing the harmonic decomposition method noise is eliminated from the audio signal. After noise removal, the features are withdrawn from the signal using discrete wavelet transform, only the requisite features are selected by the fuzzy inference system (FIS). The optimum version of these features is used for training the ANN classifier. The optimization is conducted using the Cuckoo Search (CS) optimization algorithm. For both ANN and fuzzy-based techniques, the MATLAB 2013a platform is used for implementation and the outcome obtained from both systems is compared and is shown below. The functioning of an ASR classifier has been measured by different performance metrics demonstrated in table 8.1.

Table 8.1 shows a comparison based on different parameters between Isolated Word ASR and Fuzzy based ASR.

Table 8.1 Comparison of Isolated Word Recognition and Fuzzy based ASR

Performance Metric	Isolated ASR [145]	Fuzzy based ASR [146]
Sensitivity	50%	95%
Specificity	74%	50%
False Positive rate (FPR)	26%	50%
Recognition accuracy	62%	95%
Word error rate	65%	5%

Smart systems or machinery with the capability to identify and comprehend human audio require to go through loads of enhancements to create significant and intuitive interaction between humans and machines. Regardless of excessive development in the arena of SR, still there exists an enormous difference in the recognition actions of a human and a machine.

The conclusion deduced from the performance comparison of both techniques infers that the Fuzzy based ASR with CS-ANN has attained an accuracy of 95% with an error rate of just 5% in the recognition task. When compared to the IWR system using ANN with BP algorithm the achieved accuracy is far better.

8.3 OUTCOMES OF THE RESEARCH

The objectives of the research were effectively met, and the following outcomes were achieved.

- Isolated words in the form of speech were effectively converted to the corresponding text. The ANN classifier works best with Back Propagation and gives a good performance in terms of recognition accuracy, specificity, and classifier performance. Effective noise removal techniques like wiener filtering, spectral subtraction and windowing are used.
- The approach of using a hybrid Artificial Bee Colony and Particle Swarm Optimization (ABC-PSO) for optimal feature extraction and selection of extracted features proves to increase the performance of ASR. The analogue voice signals are converted to digital voice signals for further processes like sampling,

windowing, and framing. Experimental results show that the proposed method can effectively recognize the speech data and yield better recognition accuracy, lesser word error rate and more efficiency.

- Fuzzy based feature selection and ANN optimized with the Cuckoo Search Optimization algorithm yielded a recognition accuracy of 95% with a lower word error rate of 5%. It proved to be a better technique compared to the uncertainty decoding method. The accuracy is very good when compared to the Isolated Word Recognition System using ANN with Back Propagation. The results show that the proposed system is better than the existing methods in terms of convergence rate, simplicity, and accuracy.

- An attempt to acquire a real-time news report dataset has been made to analyze the data for speech recognition and create acoustic models for studying the ASR systems. The required data is being pre-processed by min-max normalization techniques that efficiently normalize the values of the data. The normalized data is then applied using MFCC which depicts the relevant features of audio and text data. These extracted features are fed into Convolutional Neural Networks (CNN) for classification. The proposed CNN is compared with ANN which gives the efficiency in terms of accuracy, sensitivity, specificity, precision, recall, f-measure and the gaussian mean (gmean). The accuracy of the proposed CNN 96% achieved better results than ANN, 76%.

The scientific method proposed in this present study and its results will be useful for the computational implementation of intonation modelling. This research can be further used by researchers by a modification to model dynamic speech recognition systems. Speech scientists and linguistics will be able to apply the results in researching their specific areas. The knowledge of this present research could be ideally used to develop many speech-related applications.

8.4 FURTHER RESEARCH

Future researchers will have to face challenges while improving the recognition accuracy to achieve maximum levels. They will have to work upon techniques for optimizing deep learning for different applications. The SR technology has much scope for research in man-machine interaction. The methodologies explained in this thesis may be used to produce a hybrid model to improve the recognition rates

further. This thesis may help the researchers with the material and methods to develop future recognition systems. It may pave a path to producing a robust speech recognition system.

The novice researchers will have to encounter numerous difficulties while enhancing the recognition accurateness to realise supreme levels. They will have to work upon techniques for augmenting deep learning for diverse applications. The Speech Recognition technique has a lot more to explore and research in the field of human-machine communication. A lot of research still needs to be done for 'homophones.' They are the words that have the same pronunciation, but different meanings, origin, spelling and understanding. For example, words like '*ate*' or '*eight*', '*see*' or '*sea*', '*two*' or '*too*' or '*to.*'

The methods described in this thesis can be reasonably used to develop a hybrid model to enhance the scale of recognition more. This thesis might assist researchers with the knowledge and approaches to create forthcoming recognition systems. It possibly will overlay a route to develop robust speech recognition systems. The scientific method proposed in this present study and its results will be useful for the computational implementation of intonation modelling. This research can be further modified to model emotive speech in Text to speech system. The proposed intonation model can help speech scientists to do experimental research in speech technology. If one linguist or speech scientist attempts to study intonation by considering most of the dialects In Malayalam, this research will be useful for many speech application systems such as Automatic speech recognition and speaker identification system. In the case of the speaker-independent ASR system, the dialectal variation of the speech gives hurdles to the system to recognize the input speech. If it is possible to train the different dialects of speech, speech recognition will be somehow easy for the system. Intonation has an important role in the speaker recognition system to determine the authenticity of a speaker's speech. Perceptual and acoustical analysis of intonation can be used for forensic linguistics especially forensic phonetics to implement many security system applications. The knowledge of this present research could be ideally useful to develop the above-said speech applications.

REFERENCES

[1] M. Janke, M. Wand, K. Nakamura, and T. Schultz, “Further investigations on EMG-to-speech conversion,” in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 2012, pp. 365–368. doi: 10.1109/ICASSP.2012.6287892.

[2] S. Schelinski, P. Riedel, and K. von Kriegstein, “Visual abilities are important for auditory-only speech recognition: Evidence from autism spectrum disorder,” *Neuropsychologia*, vol. 65, pp. 1–11, Dec. 2014, doi: 10.1016/J.NEUROPSYCHOLOGIA.2014.09.031.

[3] M. Akbacak, L. Burget, W. Wang, and J. Hout, “Rich system combination for keyword spotting in noisy and acoustically heterogeneous audio streams,” in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, 2013, pp. 8267–8271. doi: 10.1109/ICASSP.2013.6639277.

[4] K. Konno, M. Kato, and T. Kosaka, “Speech recognition with large-scale speaker-class-based acoustic modeling,” in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Oct. 2013, pp. 1–4. doi: 10.1109/APSIPA.2013.6694112.

[5] A. Buzo, H. Cucu, L. Petrica, D. Burileanu, and C. Burileanu, “An Automatic Speech Recognition solution with speaker identification support,” 2014, pp. 1–4. doi: 10.1109/ICComm.2014.6866674.

[6] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, “A Review on Speech Recognition Technique,” *International Journal of Computer Applications*, vol. 10, no. 3, pp. 16–24, Nov. 2010, doi: 10.5120/1462-1976.

[7] B. H. Juang and L. R. Rabiner, “Automatic Speech Recognition – A Brief History of the Technology Development,” *Elsevier Encyclopedia of Language and Linguistics*, pp. 1–24, 2005, doi: 10.1016/B0-08-044854-2/00906-8.

[8] M. Arif and R. Anwar, “A New Technique for Voice Recognition in Bangla Human Gait Analysis View project Sitting Behaviour View project,” 2004. [Online]. Available: https://www.researchgate.net/publication/283498023

[9] H. Krim and M. Viberg, “Two Decades of Array Signal Processing Research: The Parametric Approach,” *Signal Processing Magazine, IEEE*, vol. 13, pp. 67–94, 1996, doi: 10.1109/79.526899.

[10] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, “A Unified Framework of HMM Adaptation with Joint Compensation of Additive and Convolutive Distortions,” *Computer. Speech Lang.*, vol. 23, no. 3, pp. 389–405, Jul. 2009, doi: 10.1016/j.csl.2009.02.001.

[11] M. A. Anusuya and S. K. Katti, “Speech recognition by machine: A review,” *International Journal of Computer Science and Information Security*, vol. 6, no. 3, pp. 181–205, 2009, doi: 10.1109/PROC.1976.10158.

[12] O. Viikki and K. Laurila, “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” *Speech Communication*, vol. 25, pp. 133–147, 1998, doi: 10.1016/S0167-6393(98)00033-8.

[13] M. Biot, “Speech Communication—Vol. 1, No. 1,” *The Journal of the Acoustical Society of America*, vol. 72, p. 291, Jul. 1982, doi: 10.1121/1.387966.

[14] O. Kalinli, M. L. Seltzer, and A. Acero, “Noise adaptive training using a vector taylor series approach for noise robust automatic speech recognition,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 3825–3828. doi: 10.1109/ICASSP.2009.4960461.

[15] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, Apr. 1995, doi: 10.1006/CSLA.1995.0010.

[16] J.-. Gauvain and Chin-Hui Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994, doi: 10.1109/89.279278.

[17] M. Sah, H. Salam, D. Mohamad, S. Salleh, M. Salam, and S. Salleh, “Malay Isolated Speech Recognition Using Neural Network: A Work in Finding Number of Hidden Nodes,” 2011. [Online]. Available: https://www.researchgate.net/publication/220413848

[18] S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000, doi: 10.1109/6046.865479.

[19] F. Ehsani and E. Knodt, “Speech Technology in Computer-Aided Language Learning: Strengths and Limitations of a New CALL Paradigm,” 2014. [Online]. Available: http://llt.msu.edu/vol2num1/article3/

[20] K. H. Hyun, E. H. Kim, and Y. K. Kwak, “Emotional Feature Extraction Based on Phoneme Information for Speech Emotion Recognition,” in *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*, Aug. 2007, pp. 802–806. doi: 10.1109/ROMAN.2007.4415195.

[21] O. Kalinli, M. L. Seltzer, J. Droppo, and A. Acero, “Noise Adaptive Training for Robust Automatic Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1889–1901, Nov. 2010, doi: 10.1109/TASL.2010.2040522.

[22] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979, doi: 10.1109/TASSP.1979.1163209.

[23] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, “A minimum-mean-square-error noise reduction algorithm on Mel-frequency cepstra for robust speech recognition,” in *International Conference on acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988*, 2008, pp. 4041–4044. doi: 10.1109/ICASSP.2008.4518541.

[24] li Deng, A. Acero, M. Plumpe, and X. Huang, “Large-vocabulary speech recognition under adverse acoustic environments.,” in *Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH 2000*, 2000, pp. 806–809.

[25] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, “The PASCAL CHiME speech separation and recognition challenge,” *Computer Speech & Language*, vol. 27, pp. 621–633, 2013, doi: 10.1016/j.csl.2012.10.004.

[26] C. Goh and K. Leon, “Robust Computer Voice Recognition Using Improved MFCC Algorithm,” in *2009 International Conference on New Trends in Information and Service Science*, Jun. 2009, pp. 835–840. doi: 10.1109/NISS.2009.12.

[27] W. Kurschl and R. Prokop, “Development issues for speech-enabled mobile applications. KeYmaera X: An aXiomatic Theorem Prover for Hybrid Systems View project,” 2007. [Online]. Available: https://www.researchgate.net/publication/221232098

[28] G. Hinton *et al.*, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.

[29] S. M. Siniscalchi, D. Yu, L. Deng, and C.-H. Lee, “Exploiting deep neural networks for detection-based speech recognition,” *Neurocomputing*, vol. 106, pp. 148–157, Apr. 2013, doi: 10.1016/J.NEUCOM.2012.11.008.

[30] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Jan. 2012, doi: 10.1109/TASL.2011.2134090.

[31] S. Molau, M. Pitz, R. Schluter, and H. Ney, “Computing Mel-frequency cepstral coefficients on the power spectrum,” in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, May 2001, vol. 1, pp. 73–76 vol.1. doi: 10.1109/ICASSP.2001.940770.

[32] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, and W.-Y. Liao, “Combining Acoustic Features for Improved Emotion Recognition in Mandarin Speech,” Springer, Berlin, Heidelberg, 2005, pp. 279–285. doi: 10.1007/11573548_36.

[33] L. Rabiner and B.-H. Juang, “An introduction to hidden Markov models,” *ASSP Magazine, IEEE*, vol. 3, no. January, p. Appendix 3A, 1986, doi: 10.1109/MASSP.1986.1165342.

[34] W. Liu, Z. Wang, X. Liu, Y. Liu, and F. E. Alsaadi, “A survey of deep neural network architectures and their applications,” *Neurocomputing*, vol. 234, pp. 11–26, 2016, doi: 10.1016/j.neucom.2016.12.038.

[35] S. Mendiratta, N. Turk, and D. Bansal, “Recognition of Human Emotional States during Automatic Speech Recognition,” *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 4, no. 4, pp. 301–307, 2016.

[36] J. Padmanabhan and M. J. Johnson Premkumar, “Machine Learning in Automatic Speech Recognition: A Survey,” *IETE Technical Review*, vol. 32, no. 4, pp. 240–251, Jul. 2015, doi: 10.1080/02564602.2015.1010611.

[37] L. Deng *et al.*, “Recent advances in deep learning for speech research at Microsoft,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8604–8608. doi: 10.1109/ICASSP.2013.6639345.

[38] A. Graves and N. Jaitly, “Towards End-To-End Speech Recognition with Recurrent Neural Networks,” in *Proceedings of the 31st International Conference on Machine Learning*, 2014, vol. 32, no. 2, pp. 1764–1772.

[39] J. Gonzalez-Dominguez, I. Lopez-Moreno, P. J. Moreno, and J. Gonzalez-Rodriguez, “Frame-by-frame language identification in short utterances using deep neural networks,” *Neural Networks*, vol. 64, pp. 49–58, Apr. 2015, doi: 10.1016/J.NEUNET.2014.08.006.

[40] D. Chen and B. Mak, “Multi-task Learning of Deep Neural Networks for Low-resource Speech Recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2015, doi: 10.1109/TASLP.2015.2422573.

[41] Y. Miao, H. Zhang, and F. Metze, “Speaker Adaptive Training of Deep Neural Network Acoustic Models Using I-Vectors,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1938–1949, Nov. 2015, doi: 10.1109/TASLP.2015.2457612.

[42] M. Henderson, B. Thomson, and S. Young, “Deep Neural Network Approach for the Dialog State Tracking Challenge,” Association for Computational Linguistics, 2013.

[43] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7304–7308. doi: 10.1109/ICASSP.2013.6639081.

[44] M. K. Mustafa, T. Allen, and K. Appiah, “A comparative review of dynamic neural networks and hidden Markov model methods for mobile on-device speech recognition,” *Neural Computing and Applications*, vol. 31, no. 2, pp. 891–899, Feb. 2019, doi: 10.1007/s00521-017-3028-2.

[45] V. Mitra, H. Franco, M. Graciarena, and D. Vergyri, “Medium-duration modulation cepstral feature for robust speech recognition,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 1749–1753. doi: 10.1109/ICASSP.2014.6853898.

[46] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, “Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 4277–4280. doi: 10.1109/ICASSP.2012.6288864.

[47] Z. Wu, T. Kinnunen, E. S. Chng, H. Li, and E. Ambikairajah, “A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case,” in *Proceedings of the*

2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, Dec. 2012, pp. 1–5.

[48] V. Khanagha, K. Daoudi, O. Pont, and H. Yahia, "Phonetic segmentation of speech signal using local singularity analysis," *Digital Signal Processing*, vol. 35, pp. 86–94, Dec. 2014, doi: 10.1016/j.dsp.2014.08.002.

[49] H. Huang and J. Lee, "A New Variable Step-Size NLMS Algorithm and Its Performance Analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 2055–2060, Apr. 2012, doi: 10.1109/TSP.2011.2181505.

[50] J. Amezquita-Sanchez and H. Adeli, "Signal Processing Techniques for Vibration-Based Health Monitoring of Smart Structures," *Archives of Computational Methods in Engineering*, vol. 23, no. 1, pp. 1–15, 2014, doi: 10.1007/s11831-014-9135-7.

[51] Z. Liu, F. Zhang, J. Wang, H. Wang, and J. Huang, "Authentication and recovery algorithm for speech signal based on digital watermarking," *Signal Processing*, vol. 123, pp. 157–166, 2015, doi: 10.1016/j.sigpro.2015.10.023.

[52] M. Suman, H. Khan, M. M. Latha, and D. Aruna Kumari, "Speech Enhancement and Recognition of Compressed Speech Signal in Noisy Reverberant Conditions," 2012, pp. 379–386. doi: 10.1007/978-3-642-27443-5_43.

[53] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for robust speech recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 4101–4104. doi: 10.1109/ICASSP.2012.6288820.

[54] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2494–2498. doi: 10.1109/ICASSP.2014.6854049.

[55] B. Li and K. C. Sim, "A Spectral Masking Approach to Noise-Robust Speech Recognition Using Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1296–1305, Aug. 2014, doi: 10.1109/TASLP.2014.2329237.

[56] Q. F. Tan and S. S. Narayanan, "Novel Variations of Group Sparse Regularization Techniques with Applications to Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1337–1346, May 2012, doi: 10.1109/TASL.2011.2178596.

[57] J. Qi, D. Wang, Y. Jiang, and R. Liu, "Auditory features based on Gammatone filters for robust speech recognition," in *2013 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2013, pp. 305–308. doi: 10.1109/ISCAS.2013.6571843.

[58] A. Sanchis, A. Juan, and E. Vidal, "A Word-Based Naïve Bayes Classifier for Confidence Estimation in Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 565–574, Feb. 2012, doi: 10.1109/TASL.2011.2162403.

[59] E. Loweimi, S. M. Ahadi, and T. Drugman, “A new phase-based feature representation for robust speech recognition,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7155–7159. doi: 10.1109/ICASSP.2013.6639051.

[60] P. Zhou, C. Liu, Q. Liu, L. Dai, and H. Jiang, “A cluster-based multiple deep neural networks method for large vocabulary continuous speech recognition,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6650–6654. doi: 10.1109/ICASSP.2013.6638948.

[61] M. Delcroix *et al.*, “Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds,” *Computer Speech & Language*, vol. 27, no. 3, pp. 851–873, May 2013, doi: 10.1016/J.CSL.2012.07.006.

[62] S. Xue, H. Jiang, L. Dai, and Q. Liu, “Speaker Adaptation of Hybrid NN/HMM Model for Speech Recognition Based on Singular Value Decomposition,” *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 175–185, Feb. 2016, doi: 10.1007/s11265-015-1012-6.

[63] L. Deng, D. Yu, and J. Platt, “Scalable stacking and learning for building deep architectures,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 2133–2136. doi: 10.1109/ICASSP.2012.6288333.

[64] S. Keronen, H. Kallasjoki, U. Remes, G. Brown, J. Gemmeke, and K. Palomäki, “Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment,” *Computer Speech & Language*, vol. 27, no. 3, pp. 2219–2231, 2012, doi: 10.1016/j.csl.2012.06.005.

[65] C. Yeh, A. Heidel, H. Lee, and L. Lee, “Recognition of highly imbalanced code-mixed bilingual speech with frame-level language detection based on blurred posteriorgram,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 4873–4876. doi: 10.1109/ICASSP.2012.6289011.

[66] L. Sun and L. Lee, “Modulation Spectrum Equalization for Improved Robust Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 828–843, Mar. 2012, doi: 10.1109/TASL.2011.2166544.

[67] P. Smaragdis and B. Raj, “The Markov selection model for concurrent speech recognition,” *Neurocomputing*, vol. 80, pp. 64–72, Mar. 2012, doi: 10.1016/J.NEUCOM.2011.09.014.

[68] G. Muhammad *et al.*, “Spectro-temporal directional derivative based automatic speech recognition for a serious game scenario,” *Multimedia Tools and Applications*, vol. 74, no. 14, pp. 5313–5327, Jul. 2015, doi: 10.1007/s11042-014-1973-7.

[69] S. J. Rennie, P. Fousek, and P. L. Dognin, “Factorial Hidden Restricted Boltzmann Machines for noise robust speech recognition,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 4297–4300. doi: 10.1109/ICASSP.2012.6288869.

[70] M. J. Alam, P. Kenny, and D. O'Shaughnessy, "Speech recognition using regularized minimum variance distortionless response spectrum estimation-based cepstral features," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8071–8075. doi: 10.1109/ICASSP.2013.6639237.

[71] H.-N. Ting, B.-F. Yong, and S. M. Mirhassani, "Self-Adjustable Neural Network for speech recognition," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 9, pp. 2022–2027, Oct. 2013, doi: 10.1016/J.ENGAPPAI.2013.06.004.

[72] Y. Kashiwagi, D. Saito, N. Minematsu, and K. Hirose, "Discriminative piecewise linear transformation based on deep learning for noise robust automatic speech recognition," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2013, pp. 350–355. doi: 10.1109/ASRU.2013.6707755.

[73] P. Dighe, G. Luyet, A. Asaei, and H. Bourlard, "Exploiting low-dimensional structures to enhance DNN based acoustic modeling in speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 5690–5694. doi: 10.1109/ICASSP.2016.7472767.

[74] Y. Li and P. Fung, "Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7368–7372. doi: 10.1109/ICASSP.2013.6639094.

[75] X. Xiao, E. S. Chng, and H. Li, "Joint spectral and temporal normalization of features for robust recognition of noisy and reverberated speech," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 4325–4328. doi: 10.1109/ICASSP.2012.6288876.

[76] X. Zhang, X. Liu, and Z. J. Wang, "Evaluation of a set of new ORF kernel functions of SVM for speech recognition," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 10, pp. 2574–2580, Nov. 2013, doi: 10.1016/J.ENGAPPAI.2013.04.008.

[77] O. Dehzangi, B. Ma, E. S. Chng, and H. Li, "Discriminative feature extraction for speech recognition using continuous output codes," *Pattern Recognition Letters*, vol. 33, no. 13, pp. 1703–1709, Oct. 2012, doi: 10.1016/J.PATREC.2012.05.012.

[78] D. Imseng, H. Bourlard, and P. N. Garner, "Using KL-divergence and multilingual information to improve ASR for under-resourced languages," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 4869–4872. doi: 10.1109/ICASSP.2012.6289010.

[79] Y. Zhang, D. Yu, M. L. Seltzer, and J. Droppo, "Speech recognition with prediction-adaptation-correction recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5004–5008. doi: 10.1109/ICASSP.2015.7178923.

[80] Y. Zouhir and K. Ouni, “A bio-inspired feature extraction for robust speech recognition,” *SpringerPlus*, vol. 3, no. 1, p. 651, Dec. 2014, doi: 10.1186/2193-1801-3-651.

[81] A. A. Nugraha, K. Yamamoto, and S. Nakagawa, “Single-channel dereverberation by feature mapping using cascade neural networks for robust distant speaker identification and speech recognition,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, p. 13, Dec. 2014, doi: 10.1186/1687-4722-2014-13.

[82] S. Zhao *et al.*, “Robust speech recognition using beamforming with adaptive microphone gains and multichannel noise reduction,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2015, pp. 460–467. doi: 10.1109/ASRU.2015.7404831.

[83] N. Esfandian, F. Razzazi, and A. Behrad, “A clustering-based feature selection method in spectro-temporal domain for speech recognition,” *Engineering Applications of Artificial Intelligence*, vol. 25, no. 6, pp. 1194–1202, Sep. 2012, doi: 10.1016/J.ENGAPPAI.2012.04.004.

[84] A. Kurakin, Z. Zhang, and Z. Liu, “A real time system for dynamic hand gesture recognition with a depth sensor,” in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Aug. 2012, pp. 1975–1979.

[85] C. Poonkuzhali, R. Karthiprakash, Dr S. Valarmathy, and M. Kalamani, “AN APPROACH TO FEATURE SELECTION ALGORITHM BASED ON ANT COLONY OPTIMIZATION FOR AUTOMATIC SPEECH RECOGNITION.” 2013.

[86] P. N. Petkov, G. E. Henter, and W. B. Kleijn, “Maximizing phoneme recognition accuracy for enhanced speech intelligibility in noise,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 1035–1045, 2013, doi: 10.1109/TASL.2013.2244089.

[87] S. JOSHI and A. N. Cheeran, “MATLAB Based Back-Propagation Neural Network for Automatic Speech Recognition,” *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 3, no. 7, pp. 10498–10504, 2014, doi: 10.15662/ijareeie.2014.0307016.

[88] D. Başkent, C. Eiler, and B. Edwards, “Phonemic restoration with hearing-impaired listeners with mild to moderate sensorineural hearing loss,” *Hearing Research*, vol. 260, no. 1–2, pp. 54–62, 2010, doi: 10.1121/1.2933722.

[89] X. Liu, M. J. F. Gales, and P. C. Woodland, “Language model cross adaptation for LVCSR system combination,” *Computer Speech & Language*, vol. 27, no. 4, pp. 928–942, Jun. 2013, doi: 10.1016/J.CSL.2012.07.010.

[90] W. Kim and R. M. Stern, “Mask classification for missing-feature reconstruction for robust speech recognition in unknown background noise,” *Speech Communication*, vol. 53, no. 1, pp. 1–11, Jan. 2011, doi: 10.1016/j.specom.2010.08.005.

[91] Shi-Xiong Zhang and M. J. F. Gales, “Structured SVMs for Automatic Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 544–555, Mar. 2013, doi: 10.1109/TASL.2012.2227734.

[92] D. Wu, Y. Yin, and H. Jiang, “Large-Margin Estimation of Hidden Markov Models with Second-Order Cone Programming for Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1652–1664, Aug. 2011, doi: 10.1109/TASL.2010.2096213.

[93] N. Hirayama, K. Yoshino, K. Itoyama, S. Mori, and H. G. Okuno, “Automatic Speech Recognition for Mixed Dialect Utterances by Mixing Dialect Language Models,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 373–382, Feb. 2015, doi: 10.1109/TASLP.2014.2387414.

[94] D. Gharavian, M. Sheikhan, and F. Ashoftedel, “Emotion recognition improvement using normalized formant supplementary features by hybrid of DTW-MLP-GMM model,” *Neural Computing and Applications*, vol. 22, no. 6, pp. 1181–1191, May 2013, doi: 10.1007/s00521-012-0884-7.

[95] S. Haque, R. Togneri, and A. Zaknich, “An Auditory Motivated Asymmetric Compression Technique for Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2111–2124, Sep. 2011, doi: 10.1109/TASL.2011.2112646.

[96] H. Kaya, T. Ozkaptan, A. A. Salah, and F. Gurgen, “Random Discriminative Projection Based Feature Selection with Application to Conflict Recognition,” *IEEE Signal Processing Letters*, vol. 22, no. 6, pp. 671–675, Jun. 2015, doi: 10.1109/LSP.2014.2365393.

[97] S. Cumani and P. Laface, “Large-Scale Training of Pairwise Support Vector Machines for Speaker Recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 11, pp. 1590–1600, Nov. 2014, doi: 10.1109/TASLP.2014.2341914.

[98] S. Chatterjee and W. B. Kleijn, “Auditory Model-Based Design and Optimization of Feature Vectors for Automatic Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1813–1825, Aug. 2011, doi: 10.1109/TASL.2010.2101597.

[99] E. Dikici, M. Semerci, M. Saraclar, and E. Alpaydin, “Classification and Ranking Approaches to Discriminative Language Modeling for ASR,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 291–300, Feb. 2013, doi: 10.1109/TASL.2012.2221461.

[100] S.-T. Pan and X.-Y. Li, “An FPGA-Based Embedded Robust Speech Recognition System Designed by Combining Empirical Mode Decomposition and a Genetic Algorithm,” *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 9, pp. 2560–2572, Sep. 2012, doi: 10.1109/TIM.2012.2190344.

[101] Y. Goh, R. Paramesran, and S. s Jamuar, “Robust speech recognition using harmonic features,” *Signal Processing, IET*, vol. 8, pp. 167–175, 2014, doi: 10.1049/iet-spr.2013.0094.

[102] H. Le, I. Oparin, A. Allauzen, J. Gauvain, and F. Yvon, “Structured Output Layer Neural Network Language Models for Speech Recognition,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1. pp. 197–206, 2013. doi: 10.1109/TASL.2012.2215599.

[103] O. Chia Ai, M. Hariharan, S. Yaacob, and L. Sin Chee, “Classification of speech dysfluencies with MFCC and LPCC features,” *Expert Systems with Applications*, vol. 39, no. 2, pp. 2157–2165, Feb. 2012, doi: 10.1016/j.eswa.2011.07.065.

[104] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, “Emotion recognition from speech using global and local prosodic features,” *International Journal of Speech Technology*, vol. 16, no. 2, pp. 143–160, Jun. 2013, doi: 10.1007/s10772-012-9172-2.

[105] M. S. Hawley *et al.*, “A voice-input voice-output communication aid for people with severe speech impairment,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 21, no. 1, pp. 23–31, 2013, doi: 10.1109/TNSRE.2012.2209678.

[106] Y. Shao and C. Chang, “Bayesian Separation with Sparsity Promotion in Perceptual Wavelet Domain for Speech Enhancement and Hybrid Speech Recognition,” *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 41, no. 2, pp. 284–293, Mar. 2011, doi: 10.1109/TSMCA.2010.2069094.

[107] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, “Silent Speech Recognition as an Alternative Communication Device for Persons with Laryngectomy,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2386–2398, Dec. 2017, doi: 10.1109/TASLP.2017.2740000.

[108] B. Wu *et al.*, “An End-to-End Deep Learning Approach to Simultaneous Speech Dereverberation and Acoustic Modeling for Robust Speech Recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1289–1300, Dec. 2017, doi: 10.1109/JSTSP.2017.2756439.

[109] D. Baby, T. Virtanen, J. F. Gemmeke, and H. Van hamme, “Coupled Dictionaries for Exemplar-Based Speech Enhancement and Automatic Speech Recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1788–1799, Nov. 2015, doi: 10.1109/TASLP.2015.2450491.

[110] S. Chandrakala and N. Rajeswari, “Representation Learning-Based Speech Assistive System for Persons with Dysarthria,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 9, pp. 1510–1517, 2017, doi: 10.1109/TNSRE.2016.2638830.

[111] S. R. Shahamiri and S. S. B. Salim, “A Multi-Views Multi-Learners Approach Towards Dysarthric Speech Recognition Using Multi-Nets Artificial Neural Networks,” *IEEE*

Transactions on Neural Systems and Rehabilitation Engineering, vol. 22, no. 5, pp. 1053–1063, 2014, doi: 10.1109/TNSRE.2014.2309336.

[112] V. Mitra, H. Nam, C. Y. Espy-Wilson, E. Saltzman, and L. Goldstein, "Articulatory Information for Noise Robust Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1913–1924, 2011, doi: 10.1109/TASL.2010.2103058.

[113] R. Sahraeian and D. Van Compernolle, "Crosslingual and Multilingual Speech Recognition Based on the Speech Manifold," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2301–2312, Dec. 2017, doi: 10.1109/TASLP.2017.2751747.

[114] Đ. T. Grozdić and S. T. Jovičić, "Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2313–2322, Dec. 2017, doi: 10.1109/TASLP.2017.2738559.

[115] J. Ming and D. Crookes, "Speech Enhancement Based on Full-Sentence Correlation and Clean Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 531–543, Mar. 2017, doi: 10.1109/TASLP.2017.2651406.

[116] J. Lee, S. Park, I. Hong, and H. Yoo, "An Energy-Efficient Speech-Extraction Processor for Robust User Speech Recognition in Mobile Head-Mounted Display Systems," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 64, no. 4, pp. 457–461, Apr. 2017, doi: 10.1109/TCSII.2016.2571902.

[117] S. M. Siniscalchi, D. Yu, L. Deng, and C. Lee, "Speech Recognition Using Long-Span Temporal Patterns in a Deep Network Model," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 201–204, Mar. 2013, doi: 10.1109/LSP.2013.2237901.

[118] O. A. Bapat, R. M. Fastow, and J. Olson, "Acoustic coprocessor for hmm based embedded speech recognition systems," *IEEE Transactions on Consumer Electronics*, vol. 59, no. 3, pp. 629–633, Aug. 2013, doi: 10.1109/TCE.2013.6626249.

[119] M. Kim, Y. Kim, J. Yoo, J. Wang, and H. Kim, "Regularized Speaker Adaptation of KL-HMM for Dysarthric Speech Recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 9, pp. 1581–1591, 2017, doi: 10.1109/TNSRE.2017.2681691.

[120] H. Hermansky, "Multistream Recognition of Speech: Dealing with Unknown Unknowns," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1076–1088, May 2013, doi: 10.1109/JPROC.2012.2236871.

[121] Y. Zhang, P. Li, Y. Jin, and Y. Choe, "A Digital Liquid State Machine with Biologically Inspired Learning and Its Application to Speech Recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 11, pp. 2635–2649, Nov. 2015, doi: 10.1109/TNNLS.2015.2388544.

[122] P. Mowlaee *et al.*, “A Joint Approach for Single-Channel Speaker Identification and Speech Separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2586–2601, Nov. 2012, doi: 10.1109/TASL.2012.2208627.

[123] M. J. Reale, P. Liu, L. Yin, and S. Canavan, “Art Critic: Multi signal Vision and Speech Interaction System in a Gaming Context,” *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1546–1559, Dec. 2013, doi: 10.1109/TCYB.2013.2271606.

[124] F. Tao and C. Busso, “Gating Neural Network for Large Vocabulary Audiovisual Speech Recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1290–1302, Jul. 2018, doi: 10.1109/TASLP.2018.2815268.

[125] S. Deena, M. Hasan, M. Doulaty, O. Saz, and T. Hain, “Recurrent Neural Network Language Model Adaptation for Multi-Genre Broadcast Speech Recognition and Alignment,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 572–582, Mar. 2019, doi: 10.1109/TASLP.2018.2888814.

[126] A. H. Abdelaziz, “Comparing Fusion Models for DNN-Based Audiovisual Continuous Speech Recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 475–484, Mar. 2018, doi: 10.1109/TASLP.2017.2783545.

[127] W. Yoon and K. Park, “Building robust emotion recognition system on heterogeneous speech databases,” in *2011 IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2011, pp. 825–826. doi: 10.1109/ICCE.2011.5722886.

[128] E. Trentin and M. Gori, “A survey of hybrid ANN/HMM models for automatic speech recognition,” *Neurocomputing*, vol. 37, no. 1–4, pp. 91–126, Apr. 2001, doi: 10.1016/S0925-2312(00)00308-8.

[129] S. Furui, “50 Years of Progress in Speech and Speaker Recognition Research,” 2005.

[130] B. H. Juang and Tsuhan Chen, “The past, present, and future of speech processing,” *IEEE Signal Processing Magazine*, vol. 15, no. 3, pp. 24–48, May 1998, doi: 10.1109/79.671130.

[131] M. Benzeghiba *et al.*, “Automatic speech recognition and speech variability: A review,” *Speech Communication*, vol. 49, no. 10–11, pp. 763–786, Oct. 2007, doi: 10.1016/J.SPECOM.2007.02.006.

[132] A. Ganapathiraju, J. E. Hamaker, and J. Picone, “Applications of Support Vector Machines to Speech Recognition,” *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2348–2355, 2004.

[133] M. Siafarikas, T. Ganchev, and N. Fakotakis, “Wavelet Packet Based Speaker Verification,” in *ODYSSEY04 -- The Speaker and Language Recognition Workshop*, 2004, pp. 257–264.

[134] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, Mar. 2011, doi: 10.1016/J.PATCOG.2010.09.020.

[135] S. Furui, “History and Development of Speech Recognition,” in *Speech Technology*, Boston, MA: Springer US, 2010, pp. 1–18. doi: 10.1007/978-0-387-73819-2_1.

[136] B. Yang and M. Lugger, “Emotion recognition from speech signals using new harmony features,” *Signal Processing*, vol. 90, no. 5, pp. 1415–1423, May 2010, doi: 10.1016/j.sigpro.2009.09.009.

[137] I. Patel and Y. Srinivas Rao, “A Frequency Spectral Feature Modeling for Hidden Markov Model-Based Automated Speech Recognition,” Springer, Berlin, Heidelberg, 2010, pp. 134–143. doi: 10.1007/978-3-642-14493-6_15.

[138] J. Manikandan, B. Venkataramani, K. Girish, H. Karthic, and V. Siddharth, “Hardware Implementation of Real-Time Speech Recognition System Using TMS320C6713 DSP,” in *2011 24th International Conference on VLSI Design*, Jan. 2011, pp. 250–255. doi: 10.1109/VLSID.2011.12.

[139] K. S. R. Murty and B. Yegnanarayana, “Combining evidence from residual phase and MFCC features for speaker recognition,” *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, Jan. 2006, doi: 10.1109/LSP.2005.860538.

[140] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, “Recurrent neural network-based language model,” in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2010, vol. 2, pp. 1045–1048.

[141] T. Mikolov and G. Zweig, “Context-dependent recurrent neural network language model,” in *2012 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2012, pp. 234–239. doi: 10.1109/SLT.2012.6424228.

[142] T. Alumäe, “Multi-domain neural network language model,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013, pp. 2182–2186.

[143] X. Chen, Y. Wang, X. Liu, M. J. F. Gales, and P. Woodland, “Efficient GPU-based training of recurrent neural network language models using spliced sentence bunch,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 641–645, 2014.

[144] X. Chen *et al.*, “Recurrent Neural Network Language Model Adaptation for Multi-Genre Broadcast Speech Recognition,” 2015.

[145] S. Mendiratta, N. Turk, and D. Bansal, “Isolated Word Recognition System for Speech to Text Conversion Using ANN,” *The official Journal of Institute of Integrative Omics and Biotechnology*, vol. 7, no. 11, pp. 78–91, 2016, doi: 20-10-2016.

[146] S. Mendiratta, N. Turk, and D. Bansal, “Fuzzy based selection of DWT features for Automatic speech recognition system for man-machine interaction with CS-ANN Classifier,” The official Journal of Institute of Integrative Omics and Biotechnology, vol. 7, no. 11, pp. 222–240, 2016.

BRIEF PROFILE OF SUNANDA MENDIRATTA

Sunanda Mendiratta was born in Shri Muktsar Sahib, Punjab. Born to a fortunate family, she studied in prestigious schools. Her primary education was at Little Flower Convent School, Muktsar. Thereafter, the middle and high school education from D. A. V. Public School, Faridabad. Her Grandfather motivated her to pursue Electronics and Communication Engineering at Maharishi Markendeshwar Engineering College, Ambala. And she received her B. Tech degree from Kurukshetra University, Haryana, India in 1999. She got married in December 2001 to Rohit Mendiratta, who is a businessman. Her supportive in-laws motivated her to get admitted to Manav Rachna International University, Faridabad for her M. Tech degree in 2011. Following her passion for academics, she got admitted to the PhD program at J. C. Bose University of Science and Technology, Faridabad. Her research interests include digital filtering techniques, artificial neural networks and speech processing. She has experience of 8 years of teaching in various prestigious engineering colleges in Faridabad. She has published papers in reputed international journals. She has also presented quite a few papers at national and international conferences.

LIST OF PUBLICATIONS

LIST OF JOURNAL PAPERS

S. No.	Title of the Paper	Name of the Journal	ISSN No.	Volume & Issue	Year
1.	Recognition of Human Emotional States During Automatic Speech recognition. https://www.ijircce.com/upload/2016/etiete/52_sunanda.pdf	International Journal of Innovative Research in Computer and Communication Engineering,	ISSN(Online) : 2320-9801 ISSN(Print):2320-9798	Vol. 4, Special Issue 4, Pg. no. 301-307	August 2016
2.	Isolated Word Recognition System for Speech to Text Conversion using ANN. http://www.iioab.org/articles/IIOABJ_7.11_78-91.pdf	Institute of Integrative Omics and Applied Biotechnology Journal,	ISSN: 0976-3104	Vol. 7, No. 11, Pg. no.78-91	October 2016
3.	Fuzzy Based Selection of DWT Features for Automatic speech recognition System for Man-machine Interaction with CS-ANN Classifier. http://www.iioab.org/articles/IIOABJ_7.11_222-240.pdf	Institute of Integrative Omics and Applied Biotechnology Journal,	ISSN: 0976-3104	Vol. 7, No. 11, Pg. No. 222-240	October 2016
4.	A Robust Isolated Automatic Speech Recognition System using Machine Learning Techniques. https://www.ijitee.org/wp-content/uploads/papers/v8i10/J87650881019.pdf	International Journal of Innovative Technology and Exploring Engineering,	ISSN: 2278-3075	Vol. 8, No. 10, Pg. No. 2325-2331	August 2019

5.	ASR system for Isolated words using ANN with Back Propagation and Fuzzy Based DWT. https://www.ijeat.org/wp-content/uploads/papers/v8i6/F9110088619.pdf	International Journal of Engineering and Advanced Technology,	ISSN: 2249–8958	Vol. 8, Issue 6, Pg No. 4813-4819	September 2019

LIST OF INTERNATIONAL CONFERENCES

S. No.	Title of the Paper	Conference Title	Conference Location	Date of Conference
6.	Automatic Speech Recognition by Cuckoo Search Optimization Based Artificial Neural Network Classifier. https://doi.org/10.1109/ICSCTI.2015.7489533 http://ieeexplore.ieee.org/document/7489533/	IEEE International Conference on Soft Computing Techniques and Implementations (ICSCTI 2015)	MRIU, Faridabad, India	October 8-10, 2015
7.	Automatic Speech Recognition Using Optimal Selection of Features Based on Hybrid ABC-PSO. https://doi.org/10.1109/INVENTIVE.2016.7824866 http://ieeexplore.ieee.org/document/7824866/	IEEE International Conference on Inventive Computation Technologies (ICICT 2016)	Coimbatore, India	August 26-27, 2016
8.	Performance Analysis of Speech Recognition Systems for Man-Machine Interaction.	International Conference on Computational Intelligence and Data Analytics (ICCIDA-2018)	GIFT, Odisha, Bhubaneswar	October 26-27, 2018

9.	Robust Feature Extraction Model for Automatic Speech Recognition System on News Report Dataset	ICTCS2021	Jaipur, Rajasthan	Decem ber 17-18,2021

LIST OF NATIONAL CONFERENCES AND SEMINARS

S. No.	Title of the Paper	Title of Conference	Location	Date
10.	Review of Speech Recognition for Man-Machine Interaction.	National Conference on Role of Science and Technology Towards Make in India	YMCAUST, Faridabad	March 5-7, 2016
11.	Automatic Speech Recognition System with Support Vector Machine Classifier.	National Conference on New Horizons in Technology for Sustainable Energy and Environment	YMCAUST, Faridabad	March 9-10, 2017
12.	Speech recognition systems for man-machine interaction – A Review.	Emerging Trends in Life Sciences Biotechnology and Conservation Strategies	K. L. Mehta Dayanand College for Women, Faridabad	March 30, 2019

LIST OF BOOK CHAPTERS

S. No.	Title of the Paper	Name of the Book	ISBN	Year
13.	Robust Feature Extraction Model for Automatic Speech Recognition System on News Report Dataset	Springer Nature Lecture Notes in Network and Systems, Vol. 400, ICTCS2021	978-981-19-0094-5	March 2022