ARTIFICIAL NEURAL NETWORK BASED AMHARIC LANGUAGE SPEAKER RECOGNITION

GIZACHEW BELAYNEH GEBRE

A Thesis submitted to the Department of Computer Science and Engineering, School of Electrical Engineering and Computing

Presented in partial fulfillment of the requirement for the Degree of Master's in Information Systems Engineering

Office of Graduate Studies

Adama Science and Technology University

ADAMA

OCT, 2019

ARTIFICIAL NEURAL NETWORK BASED AMHARIC LANGUAGE SPEAKER RECOGNITION

GIZACHEW BELAYNEH GEBRE

ADVISOR: DR. TEKLU URGESSA

A Thesis submitted to the Department of Computer Science and Engineering, School of Electrical Engineering and Computing

Presented in partial fulfillment of the requirement for the Degree of Master's in Information Systems Engineering

Office of Graduate Studies

Adama Science and Technology University

ADAMA

OCT, 2019

Approval of Board of Examiners

We, the undersigned, members of the Board of Examiners of the final open defense by ***Gizachew Belayneh*** have read and evaluated his thesis entitled ***"Artificial Neural Network Based Amharic Language Speaker Recognition "*** and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirement of the Degree of Masters of Science in Information Systems Engineering.

_____	_____	_____
Advisor	***Signature***	***Date***
_____	_____	_____
Chair Person	***Signature***	***Date***
_____	_____	_____
Internal Examiner	***Signature***	***Date***
_____	_____	_____
External Examiner	***Signature***	***Date***

Declaration

I hereby declare that this MSc. Thesis is my original work and has not been presented for a degree in any other university or institutions, and all sources of material used for this thesis have been duly acknowledged. I certify that the ideas, designs and experimental work, results, analyses and conclusions set out in this thesis are entirely my own effort, except where otherwise specified and acknowledged.

Name: Gizachew Belayneh

Signature _____

This MSc. Thesis has been submitted for examination with my approval as thesis advisor.

Name: Dr. Teklu Urgessa

Signature _____

Submission Date: October 16, 2019

መታሰቢያነቱ

ከመፈጠሬ በፊት ለነበሩት፤ እንቅፋቴን ቀድመው ለተመቱልኝ፤ ከውልደቴ ጀምሮ እስካሁን አብረውኝ ላሉት፤ የህይወት ጀግኖቼ ለሆኑት ለአባቴ ***በላይህ ንብሬ*** ለእናቴ ***ዘርጊ ንብሩ*** እና ለታላቅ ወንድሜ ***አለማየሁ በላይህ*** ይሁንልኝ።

አርዓያዬ የሆንከውና የትምህርት ፍላጎትን በልጅነቴ የከተብከኝ አባቴ **በላዬ**፤ የእኔ መንገድ እንዲሰምር የራስህን መንገድ ቀይረህ ዋጋ ስለከፈልክልኝ አመሰግናለው። ፈጣሪም ረዳኝ እንድታፍርብኝ አላደረገኝም። ነገር ግን የአንተ ብዙ ዝቅ ማለቶች ናቸው እኔን ከፍ ያደረጉኝ። የመምህርነት ልኬትን፥ ለእውነትና ላመኑበት ነገር መኖርን፥ ቤተሰባዊ ዲሞክራትነትን ከአንተ ወስጃለው። ልጅ ከአባቱ ማግኘት ኖሮበት እኔ ከአንተ ያጣሁት ምንም ነገር የለም። ለዚህም እጅግ በጣም አመሰግናለው። **እወድሃለው**!

የመንፈስ ጥንካሬና የትዕግስት ባለቤት የሆንሽው እናቴ **እቴት**፥ በአይኔ እያየሁ ያለፍሻቸው ውጥንቅጦች ለእኔ ትዕግስትን በተግባር ያስተማርሽባቸው ጊዜያቶች ናቸው። ዩኒቨርሲቲም ስገባ ሃይል የሆኑኝ እነዚያ በፅናት ያለፍሻቸው ቁስሎች ናቸው። እናቴ ጥንካሬ፥ ትዕግስት፥ ይቅርታ፥ ወርቅ የሆነ ዝምታ፥ ቆራጥነት፥ እምነት፥ አለማሽቃበጥ እና ሌሎችም የአንቺ ባህሪታ ናቸው። የወስድኳቸው አሉ የቀሩትንም የእኔ ለማድረግ እየተለማመድኳቸው ነው ። ወድቀሽ እንድነሳ፥ ወዝሽ አልቆ የእኔ ፊት እንዲያበራ ምክንያት ስለሆንሽ እጅግ በጣም አመሰግናለው። **እወድሻለው**!

ታላቄም መንትያዬም የሆንከው የእኔ የአለም ምርጡ ታላቅ ወንድም **አሌክስ**፤ ልጅነቴ በእኩዮቼ ወይም በሚበልጡኝ ያልተደፈረ እንዲሆን መከታዬ ነበርክ፥ በህይወታችን ዝቅ ብለን ስንበር እንደሚነጋ አብረን እያቀድን ተስፋ እንዲኖረኝ አድርገሃል:: እንደኔ ታላቅ ወንድሙ መንታያው ሆኖ ያደገ ያለም አይመስለኝም:: ከፍ ስንል ደግሞ ወንድምህ ከሰው እኩል እንዲሆን ጤናህን መሥዋዕት አድርገህ ታላቅ ወንድም እንዴት ማሰብ እንዳለበት አስተምረኸኛል:: ወንድሜ ዝቅ ብለህ ከፍ ስላደረከኝ እጅግ በጣም አመሰግናለው:: **እወድሃለው**!

የወደፊቱን የሚወስን ፈጣሪ ጤናችሁን ይስጥልኝና በስርዓት እንዳመስግናችሁ የእርሱ ፍቃድ ይሁንልኝ። **አሜን**!

Acknowledgement

Foremost, I would like to thanks my omnipotent **GOD** for his special endless care that helped me to accomplish this thesis work and he granted me to have the sweetest success of life in this challenging year. Next, I am glad to express my deepest gratitude to my advisor ***Dr. Teklu Urgessa*** for his supportive assistance, productive comments, criticism, and qualified advising till the completion of this thesis work. I should strongly appreciate his patience and democratic handling. I am very proud of working with you. I could get valuable life time lesson regarding research methods and life too.

I cannot miss forever to thanks you both my heroes, አባቴ ***Belayneh Gebre*** and እናቴ ***Zergi Gebru***. I am here because you paid many painful sacrifices. I am ready to pay painful sacrifices like you did to make you happy.

My special thanks should be given to my half-me wife Enatnesh Minale (***Liyu***) for her pure love, deep encouragement, unlimited and consistent kindness. ***Liyu,*** I will always avail myself for you anywhere, anytime in any situation.

Thanks to my brothers and sisters for their encouragement and deep worrying about me. Dear my gift brothers and sisters, you are source of peace for my life. I am always happy by God for granted me to have you. And thank you uncle kifle, your external hard disk helps me to do my thesis anywhere in flexible way.

I honestly acknowledge for those I used their speech, I used your speech not because of I am senseless about you, rather you are at the top in my mind shelf.

Mr. Anwar Bilcha (*upcoming Dr.****),*** missing to thanks you is impossible, you showed me not only the right track, it was the best track too. You protected me from invisible mess.

[Dear ATE], I don't know how I thanks you. Without your support, this thesis wouldn't be real.

Finally, it is my pleasure to thank my classmates. Dear classmates, you deserve to be remembered here. You gave me a life time memorable experiences in addition to commenting my thesis work. ***Thank you!!***

Table of Contents

Declaration ii
Acknowledgement iv
List of Figures and Tables vii
Abstract xi
CHAPTER ONE: INTRODUCTION 1
1.1. Background 1
1.2. Motivation of the Study 3
1.3. Statement of the Problem 3
1.4. Research Questions 5
1.5. Objectives 5
1.5.1. General Objective 5
1.5.2. Specific Objectives 5
1.6. Scope and Limitation 5
1.7. Significant of the Study 6
1.8. Thesis Organization 7
CHAPTER TWO: LITERATURE REVIEW 8
2.1. Speech Signal 8
2.2. Speech Processing 9
2.2.1. Speech Synthesis 9
2.2.2. Speech Recognition 9
2.2.3. Speaker Recognition 10
2.2.3.1. Classification of Speaker Recognition 11
2.2.3.2. Speaker Recognition Techniques 13
2.2.3.3. Applications of Speaker Recognition 17
2.3. Artificial Neural Networks 18
2.3.1. Introduction 18
2.3.2. Elements of Artificial Neural Networks 19
2.3.3. ANN Architectures 22
2.3.3.1. Single-Layer Feedforward Networks 22
2.3.3.2. Multi-Layer Feedforward Networks 23
2.3.3.3. Recurrent Networks 24
2.3.4. Application of Artificial Neural Network 24

2.4. Amharic Language 25
2.4.1. Amharic Phonetics 26
2.4.1.1. Amharic Consonants 27
2.4.1.2. Amharic Vowels 30
2.5. Related Works 31
CHAPTER THREE: METHODODOLGY 36
3.1. Framework of Methodology 36
3.2. Research Design 37
3.3. Data Source Identification 37
3.4. Speech Collections 37
3.5. Preprocessing 38
3.5.1. Silence Removal 38
3.5.2. Noise Reduction and Normalization 39
3.5.3. Splitting of Audio Speech Signals. 40
3.6. Feature Extraction 41
3.7. Feature Modeling and Matching Technique 46
3.8. Tools 49
3.9. Evaluation Methods 49
3.10. Summary 50
CHAPTER FOUR: EXPERIMENTS, FINDINGS AND DISCUSSION 51
4.1. Experiments and Findings 51
4.1.1. Proposed Speaker Recognition Prototype 51
4.1.2. Feature Extraction 52
4.1.3. Training of Neural Network 53
4.1.4. Performance Evaluation 57
4.2. Discussions 68
CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS 71
5.1. Conclusions 71
5.2. Recommendations 73
5.3. Future Works 73
References 74
Appendixes lxxix

List of Figures and Tables

Figure 1. 1 Sample Speech Signal 8
Figure 1. 2 Process flow of the Main Components of Speaker Recognition System [16]. 11
Figure 1. 3 Classification of Speaker Recognition 12
Figure 1. 4 Block Diagram of Steps in MFCCs Feature Extraction (Adopted)..... 15
Figure 1. 5 Layers in Simple Artificial Neural Network 20
Figure 1. 6 Elements of Artificial Neural Network 22
Figure 1. 7 Single-Layer Feedforward Network..... 23
Figure 1. 8 Multi-Layer Feedforward Network 23
Figure 1. 9 Recurrent Neural Network Architecture 24
Figure 1. 10 Conceptual Framework of the Methodology..... 36
Figure 1. 11 AA.wav Original Audio 38
Figure 1. 12 AA.wav after Silence Removal 39
Figure 1. 13 AA.wav after Noise Reduction..... 39
Figure 1. 14 AA.wav after Normalization 39
Figure 1. 15 Single Frame of AA_00.wav..... 42
Figure 1. 16 Frame and Windowed Frame of AA_00.wav..... 43
Figure 1. 17 FFT of single Windowed Frame of AA_00.wav..... 44
Figure 1. 18 Mel-Filter Bank of Single Windowed Frame of AA_00.wav 45
Figure 1. 19 MFCC Feature of AA_00.wav 45
Figure 1. 20 Neural Network of Experiment One..... 48
Figure 1. 21 Architecture of Proposed Speaker Recognition Prototype..... 51
Figure 1. 22 MFCC Feature Vector Matrix of AA_00.wav 52
Figure 1. 23 Neural Network Training Tool; Training of Experiment One 54
Figure 1. 24 Performance Plot of Experiment One [Hidden Neurons=15] 55
Figure 1. 25 Performance Plot of Experiment Two [Hidden Neurons=20]..... 56
Figure 1. 26 Performance Plot of Experiment Three [Hidden Neurons=25]..... 57
Figure 1. 27 Experiment One: Confusion Matrix [Hidden Neurons=15]..... 60
Figure 1. 28 Experiment Two: Confusion Matrix [Hidden Neurons=20] 61
Figure 1. 29 Experiment Three: Confusion Matrix [Hidden Neurons=25] 62

Figure 1. 30 Average Precision, Recall, F1-score, and overall Accuracy of all Experiments 67

Table 1. 1 Amharic Consonants 28
Table 1. 2 Dictionary Definition of Terms 29
Table 1. 3 Amharic Vowels 30
Table 1. 4 Summary of Related Works 35
Table 1. 5 Training and Testing Datasets 40
Table 1. 6 Summary of Techniques & Tools and their Purpose 50
Table 1. 7 Training dataset and their corresponding class ID after feature extraction 53
Table 1. 8 Result Summary Based on Gender 63
Table 1. 9 Confusion Matrix Metrics Result of each Class of all Experiments 64
Table 1. 10 Computed Model and Classification Accuracy of all Experiments 66
Table 1. 11 Real Time Testing Recognition Results 67
Table 1. 12 Comparison of this research with other related researches 70

List of Abbreviations and Acronyms

Abbreviation / Acronyms	Stands
5G	Fifth Generation
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BDNN	Bi-Directional Neural Network
CV	Consonant-Vowel
DCT	Discrete Cosine Transform
DFNN	Deep Forward Neural Network
DFT	Discrete Fourier Transform
DTW	Dynamic Time Warping
FFNN	Feed Forward Neural Network
FFT	Fast Fourier Transform
GFCC	Gammatone Frequency Cepstral Coefficients
GMM	Gaussian Mixture Model
GUI	Graphical User Interface
HMM	Hidden Markov Model
ICA	Independent Component Analysis
KHz	Kilo Hertz
LDA	Linear Discriminate Analysis
LPC	Linear Predictive Codes
LPCC	Linear Prediction Cepstrum Coding
LSTM	Long-Short Time Memory
MATLAB	Matrix Laboratory
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multiple Layer Perceptron
MSE	Mean Squared Error
NN	Neural Network
PCA	Principal Component Analysis

PIN	Personal Identification Number
PLP	Perceptual Linear Prediction
RNN	Recurrent Neural Network
STT	Speech to Text
SVM	Support Vector Machine
TTS	Text to Speech
UMRT	Unique Mapped Real Transform
VQ	Vector Quantization

Abstract

In this artificial intelligence time, speaker recognition is the most useful biometric recognition technique. Security is a big issue that needs careful attention because of every activities have been becoming automated and internet based. For security purpose, unique features of authorized user are highly needed. Voice is one of the wonderful unique biometric features. So, developing speaker recognition based on scientific research is the most concerned issue. Nowadays, criminal activities are increasing day to day in different clever way. So, every country should have strengthen forensic investigation using such technologies. The thesis was done by inspiration of contextualizing this concept for our country. In this study, text-independent Amharic language speaker recognition model was developed using Mel-Frequency Cepstral Coefficients to extract features from preprocessed speech signals and Artificial Neural Network to model the feature vector obtained from the Mel-Frequency Cepstral Coefficients and to classify objects while testing. The researcher used 20 sampled speeches of 10 each speaker (total of 200 speech samples) for training and testing separately. By setting the number of hidden neurons to 15, 20, and 25, three different models have been developed and evaluated for accuracy. The fourth-generation high-level programming language and interactive environment MATLAB is used to conduct the overall study implementations. At the end, a very promising findings have been obtained. The study achieved better performance than other related researches which used Vector Quantization and Gaussian Mixture Model modeling techniques. Implementable result could obtain for the future by increasing number of speakers and speech samples and including the four Amharic accents.

Keywords: *Text-independent Speaker Recognition, ANN, MFCC, Feature Extraction, Amharic Language, MATLAB Neural Network Tool.*

CHAPTER ONE: INTRODUCTION

1.1. Background

Speech is a medium for human to express their thoughts during communication. A speech signal is a complex signal which is packed with several knowledge resources such as acoustic, articulatory, semantics, linguistic and many more. It is the ultimate, universal mode of human communication and it is how a man should be able to interact with computers or machines[1]. Speech interface in the user's own language is in short, an ideal means of communication as being the most natural, flexible, efficient and convenient option allowing to perform hands and eyes free tasks. The study of speech signals and the processing methods of signals is called speech processing. It is the analysis of human speech (using digital signal processing techniques). There are several aspects of speech processing, according to the focus of analysis: speech synthesis, speech recognition, speaker recognition, voice analysis, speech coding and compression, speech enhancement, speaker diarization, speaker classification, language identification, etc. ***Speaker recognition*** which is the concern of this study is the process of automatically recognizing who is speaking by using the speaker-specific information included in speech waves to verify identities being claimed by people accessing systems; that is, it enables access control of various services by voice [2]. The aim of speaker recognition is to extract, characterize and recognize the information about speaker identity.

In this digital world, speaker recognition is the most useful biometric recognition technique. Now days many organizations like bank, industries, access control systems, etc. are using this technology for providing greater security to their vast databases [3]. Applicable services include voice dialing, banking over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information, and remote access to computers. The most essential application of speaker recognition technology is as a forensics tool.

Speaker recognition can be classified into speaker identification and speaker verification. ***Speaker identification*** is the process of determining from which of the registered speakers a given utterance comes. ***Speaker verification*** is the process of accepting or rejecting the

identity claimed by a speaker. Most of the applications in which voice is used to confirm the identity of a speaker are classified as speaker verification. From a security perspective, identification is different from verification. For example, presenting your passport at border control is a verification process: the agent compares your face to the picture in the document. Conversely, a police officer comparing a sketch of an assailant against a database of previously documented criminals to find the closest match(s) is an identification process.

Speaker recognition systems fall into two categories, text-dependent, and text-independent. ***Text-dependent*** uses the same text for enrollment and testing. ***Text-independent*** uses different text for enrollment and testing. This study is dealt with text-independent speaker identification in a case of Amharic language speech.

At the highest level, all speaker recognition systems contain two main modules ***feature extraction*** and ***feature modeling and matching.*** Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature modeling and matching is modeling the extracted features and involves the actual procedure to identify the unknown speaker by comparing extracted features from the input voice with the ones from a set of known speakers.

However, the speech based systems that have been developed so far are meant to serve a specific language and are totally limited to some techno-rich countries of the world. For developing countries like Ethiopia, it is a must, to follow the outfit of those techno-rich countries in relation to such technological advancements to do not lose the opportunities provided by technologies. Based on this fact, speech engineers and language experts in various countries are making noticeable efforts to develop recognition that works for their own language. In our country, even though it is not enough 40 speech recognition and 3 speaker recognition researches had been attempted. As we have seen, there is a shortage of research regarding speaker recognition. The above three speaker recognition concerned researches used Vector Quantization and Gaussian Mixture model for modeling techniques. The goal of this study is exploring the possibility of a state of the art modeling techniques for building Amharic language speaker recognition.

1.2. Motivation of the Study

Nowadays, technology is entirely becoming 5G in which the initial development of smart devices tasks that could only be performed on a desktop or laptop computer in the past could suddenly be performed on a smart platform, such as web browsing. In 5G world, internet speed, processing capability, memory size will not be a problem. Technologies are undertaking in the consideration to replace (minimize) human role in any activity. It is a movement towards artificial intelligence. Speech technology is one of these areas that gets attention. Communicating with machine in better performance like human to human communication is the main objective of today's researcher. In addition to speech recognition and TTS, speech can be used as biometrics.

As far as the researcher tried to search published and unpublished researches which are related to speech technology, around 40 speech recognition and 3 speaker recognition researches have been done for local languages in our country. This shows, there is a dearth of research on speaker recognition. Three of speaker recognition researches have been attempted using Vector Quantization and Gaussian Mixture Modeling as modeling technique, which have no feature discriminating ability as artificial neural network for speaker recognition. Because of this, the researcher is inspired to attempt and contribute an Amharic language speaker recognition using state of the art modeling technique called artificial neural network.

1.3. Statement of the Problem

In this digital period, speaker recognition is the most useful biometric recognition technique. Today many organizations like bank, industries, access control systems, etc. are using this technology for providing greater security to their vast databases [3]. Most of the researchers have found that using biometric as the security system is the most effective and efficient solution for speaker identification and verification. Because biometric can only verify the unique characteristic, cannot be copy by anyone. Speaker recognition is one of the widest research areas where biometrics techniques can be used voice for security purpose as the identification. The most essential application of speaker recognition technology is forensic investigation. For a long time, there has been a need to be capable of identifying an individual based on their voice. For a number of years, law enforcement

agencies judges, lawyers, and detectives wanted to utilize forensic voice authentication for investigating a suspect or confirming a judgment of innocence or guilt [8].It is obvious that criminal activities are increasing day to day in different clever way. So, forensic investigations should be supported with biometric techniques like speaker recognition.

And also, speaker identification system is essential and can be preferred for attendance in governmental and non-governmental institutions such as offices, colleges, and universities to control the flow of stored private data and financial transactions. So, speaker identification is more convenient than traditional means of authentication such as password and PIN.

On the other hand, improving accuracy rate of identification is great issues and still a big challenge for researchers. Researchers are trying to enhance the accuracy of the system because of accuracy of the system is definitely a major issue. Thus, it is absolutely necessary to discover the cause of accuracy gap in speaker recognition systems. Discovering the source reason allows to increase the accuracy of speaker recognition on the previous studies.

In our country, speaker recognition has no application; for instance, the forensic investigation is not being supported by such technologies[1] . A few researches had been attempted regarding speaker recognition, but still shows the dearth of researches that can enable to transform into practical. Thus, it is unquestionable to conduct many researches as much as possible and develop robust speaker identification and recognition systems to increase safety and security.

As a result, this study aimed to address the above-stated problems and experimental development of text-independent speaker recognition system for Amharic utterances in order to discovery solutions to the next research questions and met the specific objectives.

[1] *The Academic Vice President of Ethiopian Police University College, Mr. Ayele Mulugeta on crime investigation, he said, "It can be categorized into tactical and technical types, forensic science comes with the latter, in this regard, we have gaps in Ethiopia, because of two reasons, first, we lack technology itself and there is dearth of knowledge and skills as well." Accessed on September 25, 2019 < https://www.amu.edu.et >*

1.4. Research Questions

Following are the research questions, which the researcher tried to answer in this study.

1. Has ANN based model a promising performance for speaker recognition?
2. What are internal key performance indicators in ANN based modeling?
3. What are external key performance indicators in building speaker recognition system?

1.5. Objectives

1.5.1. General Objective

The general objective of this study is exploring the possibility of using artificial neural network modeling technique for the development of text-independent Amharic language speaker recognition.

1.5.2. Specific Objectives

- Relevant Literature Review
- Data Source Identification.
- Speech collections.
- Preprocess speech signals.
- Feature Extraction of preprocessed speech signals.
- Modeling using neural network.
- Develop a prototype.
- Evaluate the performance of the model and the prototype.
- Interpret the results and draw conclusions.

1.6. Scope and Limitation

The research is limited to text-independent Amharic language speaker recognition using an artificial neural network. It is only concerned on one category of speaker recognition which is closed-set speaker identification. It doesn't include speaker verification. The research is also limited to processing only the speech data from the speaker person, it didn't

consider the speaker's health condition and emotion recognition. All speech signals are preprocessed and changed in to noise free. So, this study didn't represent the performance of the model for speech signals recorded at noisy environment. In speaker perspective, even though there are different accent of Amharic language such as Gojjam, Gondar, Shewa, and Wollo, in this study considered only native shewa accent Amharic speakers.

1.7. Significant of the Study

This study will have its own role in supporting the efforts in making technology reachable and suitable for the whole community. Many applications have been considered for speaker recognition. Today in everywhere rapidly growing technological security devices to entail the need of reliable user authentication technique to prevent identity theft and enhance regional as well as national security from foreign attack. Such kind of research conducting speaker recognition will be worthy for many sectors governmental, non-governmental as well as individuals can be benefited from speaker identification as result of the good biometric system. Speaker identification is a good biometric system so, many organizations and individuals can be benefited from it. Forensic investigation experts, courts, police commissions can also be benefited from the study for identifying individuals in criminal justice and terrorist actions. It could be able to solve problems of physically disabled people, as an assistive or ideal tool combined with speech recognition and improving the human-computer interaction interface mechanism for the local languages. This study is, therefore, a good starting towards increasing the chance to develop such a useful system. As the study is tried to use ANN approach for the first time, it will introduce a new feature modeling and matching technique attempt for Amharic language speaker recognition development and it will create a chance for the coming researchers to compare and contrast the performance of this technique with others.

1.8. Thesis Organization

This thesis is organized in five chapters and the contents of each chapter are described as follows.

The first chapter which is chapter one introduces the whole paper and is consisted of the subtopics: background of the study, statement of the problems, objectives of the research, methods, the significance of the study and finally scope and limitation of the study. Chapter Two: Literature Review, discusses the speech processing applications, definition, and classification of speaker recognition, speaker recognition approaches, techniques, and methods, linguistics behavior of Amharic. Finally, some previously done related works are presented. Chapter Three: Research Methodology, it describes the methodological procedures followed in doing this research. Chapter Four: Experiment, Results, and Discussions, it describes how the experiment is done, what are the findings and their interpretation. Chapter Five: Conclusion and Recommendations, concludes the thesis by outlining achievements, drawbacks and possible future works to improve the work carried out by this research.

CHAPTER TWO: LITERATURE REVIEW

2.1. Speech Signal

Speech is defined as the expression of thoughts and feelings by articulating sounds. Speech is the most natural, intuitive and preferred means of communication by human beings. The perceptual variability of speech exists in the form of various languages, dialects, accents, while the vocabulary of speech is growing day by day. More intricate variability at the speech signal level exists in the form of varying amplitude, duration, pitch, timbre and speaker variability [15].

Speech signals are highly non-stationary (vary in time), with a wide dynamic range of multiple frequency components in the short-time spectra. The speech signal, as it emerges from a speaker's mouth, nose, and cheeks, is a one-dimensional function (air pressure) of time. Microphones convert the fluctuating air pressure into electrical signals, voltages or currents, in which form we usually deal with speech signals in speech processing.

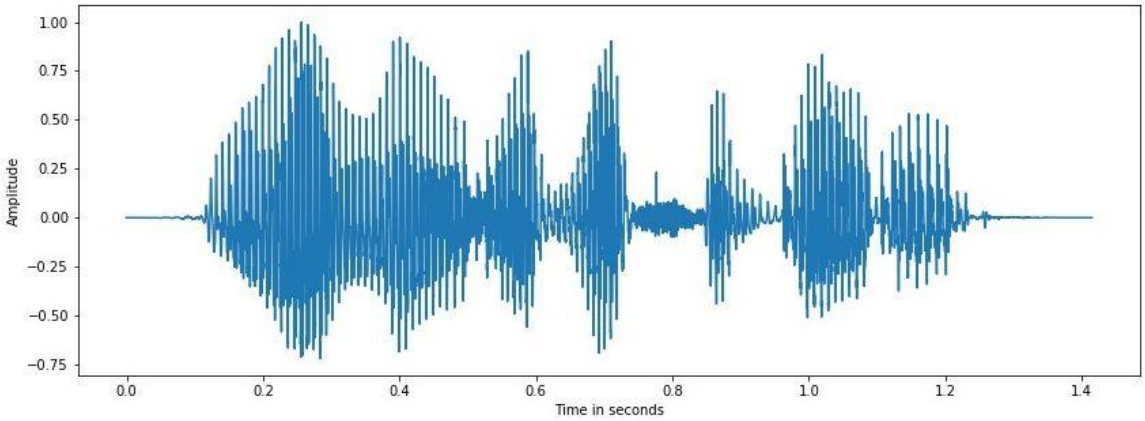


Figure 1. 1 Sample Speech Signal

2.2. Speech Processing

Speech processing is the study of speech signals and the processing methods of signals. The signals are usually processed in a digital representation, so speech processing can be regarded as a special case of digital signal processing, applied to speech signals is the extraction of the needed information from a speech signal to be digitalized and processed by the computer. There are several tasks (applications) of speech processing, among the tasks (applications); speech synthesis, speech recognition, and speaker recognition are common.

2.2.1. Speech Synthesis

Speech synthesis is the computer-generated simulation of human speech. It is used to translate written information into aural information where it is more convenient, especially for mobile applications such as voice-enabled e-mail and unified messaging. It is also used to assist the vision-impaired so that, for example, the contents of a display screen can be automatically read aloud to a blind user. Speech synthesis is the counterpart of speech or voice recognition. Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

2.2.2. Speech Recognition

Speech is the ultimate, universal mode of human communication and it is how a man should be able to interact with computers or machines [16]. Speech interface in the user's own language is in short, an ideal means of communication as being the most natural, flexible, efficient and convenient option allowing to perform hands and eyes free tasks. Speech recognition is the interdisciplinary subfield of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. It is also known as automatic speech recognition (ASR), computer speech recognition or speech to text (STT). It incorporates knowledge and research in linguistics, computer science, and electrical engineering fields. The ultimate

goal of the technology is to be able to produce a system that can recognize with 100% accuracy all words that are spoken by any person.

2.2.3. Speaker Recognition

It is the concern of this study, is the process of automatically recognizing who is speaking by using the speaker-specific information included in speech waves to verify identities being claimed by people accessing systems; that is, it enables access control of various services by voice [2]. The aim of speaker recognition is to extract, characterize and recognize the information about speaker identity.

Now days many organizations like bank, industries, access control systems, etc. are using this technology for providing greater security to their vast databases [3]. Applicable services include voice dialing, banking over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information, and remote access to computers. Another important application of speaker recognition technology is as a forensics tool.

At the highest level, all speaker recognition systems contain two main modules feature extraction and feature modeling & matching. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from the input voice with the ones from a set of known speakers. Speaker recognition is classified into speaker identification and speaker verification. Speaker identification is the process of determining from which of the registered speakers a given utterance comes. Speaker verification is the process of accepting or rejecting the identity claimed by a speaker. Figure 1.2. Shows general process of speaker recognition system.

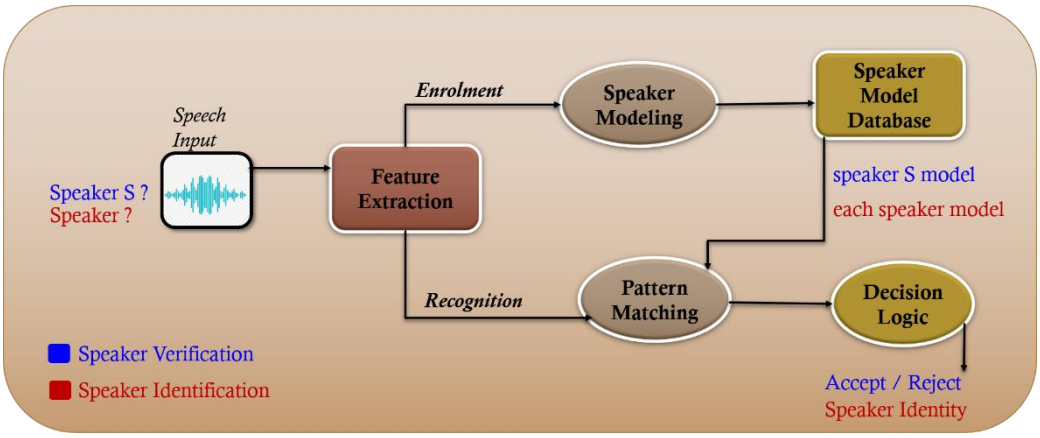


Figure 1. 2 Process flow of the Main Components of Speaker Recognition System [16].

2.2.3.1. Classification of Speaker Recognition

Speaker recognition can be classified into speaker identification and speaker verification. Speaker identification is the process of determining from which of the registered speakers a given utterance comes. Speaker verification is the process of accepting or rejecting the identity claimed by a speaker. Most of the applications in which voice is used to confirm the identity of a speaker are classified as speaker verification. Both speaker identification and speaker verification need a stored speaker database as a reference model for pre known speakers. The fundamental difference between identification and verification is the number of decision alternatives. In identification, the number of decision alternatives is equal to the size of the population, whereas in verification there are only two choices, acceptance or rejection, regardless of the population size. From a security perspective, identification is different from verification. For example, presenting your passport at border control is a verification process: the agent compares your face to the picture in the document. Conversely, a police officer comparing a sketch of an assailant against a database of previously documented criminals to find the closest match(s) is an identification process.

Also, on the basis of dependency on text, speaker recognition systems fall into two categories, text-dependent and text-independent. In text-dependent systems, the spoken text is previously known in both training and testing phases, where in text-independent systems the speaker is allowed to speak freely. The system relies on text-independent model and should figure out the distinguishing speech characteristics of vocal sounds that belong to speakers participating in that system. On the other hand, the text-dependent

system needs specific phrases to be spoken (like card number, passwords, etc.) which leads to better speaker recognition results [17]. This study is dealt with text-independent speaker identification in a case of Amharic language speech.

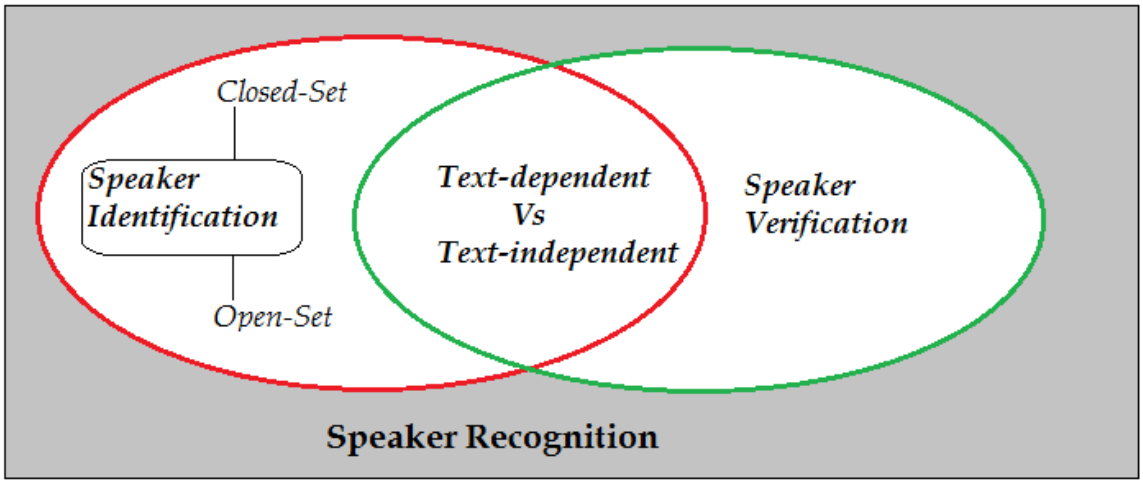


Figure 1. 3 Classification of Speaker Recognition

A. Speaker Identification

Speaker identification (present task of this study) is the process of determining from which of the registered speakers a given utterance comes. Speaker identification system is one to many matching system. In this, user needs not to provide his/her identity to the system. In the speaker identification task, a speech utterance from an unknown speaker is analyzed and compared with speech models of known speakers. The unknown speaker is identified as the speaker whose model best matches the input utterance [18]. In this case, system has to perform N comparison (N is the number of stored speaker model in database). Then, comparison with each registered model will produce a probability percentage value, on the basis of this percentage value higher probability model is identified for the speaker. From this, it is clear that speaker identification is complex than speaker verification. Speaker identification categorized in to ***closed-set*** and ***open-set*** identification. In ***Closed-Set identification***, any subject presented to the biometric system for recognition is known to be enrolled in the system; thus, no rejection is needed in principle unless the quality of the input biometric trait is too low to process. Means, the unknown voice come from a fixed set of known speakers. In ***Open-Set identification***, it is unknown whether the subject presented to the biometric system for recognition has enrolled in the system or not.

Therefore, the system needs to decide whether to reject or recognize him as one of the enrolled subject.

B. Speaker Verification

Speaker verification is the process of accepting or rejecting the identity claimed by a speaker. Speaker verification system is one to one matching system that either accepts or rejects. In verification process, firstly user need to provide his/her identity and then it is checked by the system and decisions are made accordingly that whether the claimed identity is true (accept) or false (reject) [3]. When a test utterance comes with a claim, it tests its parameters under the claimed speaker's model and calculates a similarity score. The decision is made depending on this score. If the score is below the threshold then the system will reject the test utterance, otherwise the system will accept the test utterance [19].

2.2.3.2. Speaker Recognition Techniques

Speaker recognition has two basic modules; feature extraction and feature modeling & matching. Feature extraction is the process of extracting set of features from a speech signal that represent a specific speaker. Feature matching is the actual procedure of recognition, which finds the best match between the extracted features of unknown input voice signal and the set of predefined speakers stored in database of the system [17]. There are several feature extraction and feature matching techniques. The most common techniques among them are discussed below.

A. Feature Extraction Techniques

Feature extraction is the process of extracting set of features from a speech signal that represent a specific speaker. The importance of feature extraction is due to the fact that processing all the data samples of the recorded speech signal in order to identify the speaker is a redundant process since not all the samples contain useful information [20]. Speech feature extraction is the signal processing front-end which converts the speech waveform into some useful parametric representation [21]. It is the most important part of speaker recognition since it plays an important role in separate one speech from others. It involves simplifying the number of resources required to describe a large set of data

accurately. The aim of feature extraction is to reduce the dimensionality of the signal by suppressing or ignoring unwanted information and retaining and enhancing speaker-specific information [22]. The most widely used feature extraction techniques are linear predictive coding (LPC), linear predictive cepstral coefficients (LPCC), and Mel-Frequency Cepstral Coefficients (MFCCs)

- **Mel-Frequency Cepstral Coefficients (MFCCs)**

Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition which is used in this study. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since. Mel frequency Cepstrum is a representation of a short term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel frequency cepstral coefficients are the coefficients that collectively make up the MFCC [26]. The most prevalent and dominant method used to extract spectral features is calculating Mel-Frequency Cepstral Coefficients (MFCC). MFCCs are one of the most popular feature extraction techniques used in speaker recognition based on the frequency domain using the Mel scale which is based on the human ear scale [25]. To simplify the consecutive processing of the signal, useful features must be extracted and the data should be compressed.

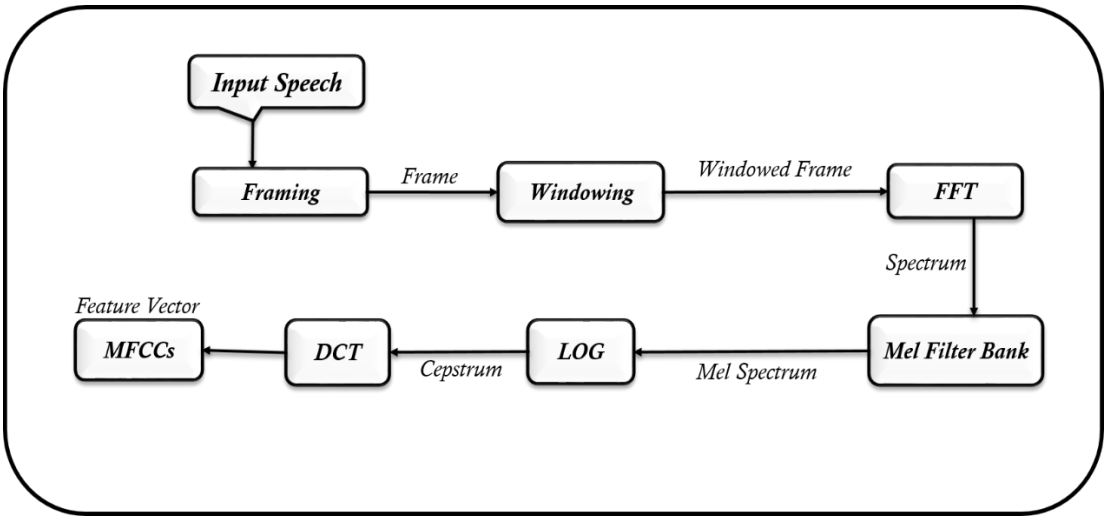


Figure 1. 4 Block Diagram of Steps in MFCCs Feature Extraction (Adopted[2])

Step1: Framing

The audio signals are continuously changing. However, in order to get a pattern, it is necessary to take a constant signal. The reason behind dividing the speech signal into frames of small duration is that the speech signal is non-stationary and its temporal characteristics change very fast. So, by taking a small frame size, we make an assumption that the speech signal will be stationary and its characteristics will not vary much within the frame. The frame length is important due to the tradeoff between time and frequency resolution. If it is too long it will not be able to capture local spectral properties and if too short the frequency resolution would degrade. Speech signal is divided into N samples per frame. Adjacent frames are being separated by M (M<N). Here, the speech is segmented into small duration blocks of 30ms. The segmenting duration standard is 20-40ms.

Step2: Windowing

In dictionary, window is the time period that is considered best for starting or finishing something. So, windowing means setting best starting and finishing time for each frame. It is used to minimize the signal discontinuity.

[2] *Smita B Magre et. al, A Review on Feature Extraction and Noise Reduction Technique*

In speech processing the shape of the window function is not that crucial but usually some soft window like Hanning, Hamming, triangle, half parallelogram, not with right angles.

Step3: Fast Fourier Transform

FFT is an algorithm that computes the discrete Fourier transform (DFT) of a sequence, or its inverse (IDFT). Fourier analysis converts a signal from its original domain (often time or space) to a representation in the frequency domain and vice versa. A speech signal represented in the time domain by a vector of amplitudes. So, Fourier Transform (FT) can be used to transform the amplitudes in the time domain into frequencies in the frequency domain periodic functions which contained unique information. It is used to convert each frame of N samples from time domain to frequency domain. FFT is perform to obtain magnitude of frequency response of each frame.

```
FFT= (abs (fft(frame)))
```

Step4: Mel Filter Bank Processing

The range of frequencies is very wide in FFT and voice signal does not follow the linear scale. The output of FFT is multiplied by a set of 32 triangular band-pass filter to get log energy of each triangular band-pass filter. Triangular band pass filters are used to extract the spectral envelope, which is constituted by dominant frequency components in the speech signal. The below equation is used to convert linear scale frequency into Mel scale frequency

```
F (Mel) = 2595 * log 10 (1+f/ 700)
```

Step 5: Discrete Cosine Transform

DCT expresses a finite sequence of data points in terms of a sum of cosine functions oscillating at different frequencies. It is a process to convert the log Mel spectrum into time domain. The result is called Mel Frequency Cepstrum Coefficients.

B. Feature Modeling and Matching Techniques

Feature modeling is training the speakers feature vector to the modeling technique. Speaker models are constructed from the extracted features. Feature matching is the actual procedure of recognition, which finds the best match between the extracted features of unknown input voice signal and the set of predefined speakers stored in database of the system [17]. Therefore, it is most significant to use fitting feature matching technique to obtain the accurate result. There are numerous feature modeling and matching techniques which can be used for speaker recognition. These techniques can be categorized in statistical techniques, soft-computing techniques and hybrid techniques. Statistical techniques include DTW, VQ, GMM, HMM, etc., while soft-computing techniques are ANN, SVM, and Fuzzy logic, etc. Hybrid techniques make use of both techniques [13].

2.2.3.3. Applications of Speaker Recognition

Many applications have been considered for speaker recognition. These include secure access control by voice, customizing services or information to individuals by voice, indexing or labeling speakers in recorded conversations or dialogues, surveillance, and also criminal and forensic investigations involving recorded voice samples. Access control applications include voice dialing, banking transactions over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, and remote access to computers [20].

The following discussion regarding the application of speaker recognition is according to (N. Singh et al, 2012) which is specifically studied on application of speaker recognition and discussed in better way [31].

Speaker Recognition for Authentication: Speaker recognition for authentication allows the users to identify person using their voices. A person can be identified by various characteristics like signature, fingerprints, voice, facial features, etc. This type of authentication methods known as biometric person authentication. In this case, the chance of misused of these type of identity problems are lesser as compared to the key or credit card can be stolen or lost, followed by the PIN number or password can be easily misused or forgotten. Each person has unique anatomy, physiology and learned habits that familiar

persons use in everyday life to recognize the person. This can be much more convenient than traditional means of authentication which require to carry a key with you or remember a PIN.

Speaker Recognition for Surveillance: Security agencies have several means of collecting information. One of these is electronic eavesdropping of telephone and radio conversations. As these results in high quantities of data, filter mechanisms must be applied in order to find the relevant information. One of these filters may be the recognition of target speakers that are of interest for the service.

Forensic Speaker Recognition: It is an important application of speaker recognition. If there is a speech sample that was recorded during the crime. The suspect's voice can be compared with this in order to give an indication of the similarity of the two voices. Proving the identity of a recorded voice can help to convict a criminal or discharge an innocent in court. Although this task is probably not performed by a completely automatic speaker recognition system, signal processing techniques can be used in this field nevertheless.

Security: It is the most obvious application of any biometric authentication techniques. Speaker recognition could be used in credit card transactions as an authentication method combined with some others like face recognition. Speaker recognition technology can provide transaction authentication facility or computer access control, monitoring, telephone voice authentication for long distance calling or banking access, etc.

2.3. Artificial Neural Networks

2.3.1. Introduction

The brain contains neurons which are kind of like organic switches. These can change their output state depending on the strength of their electrical or chemical input [32]. The neural network in a person’s brain is a hugely interconnected network of neurons, where the output of any given neuron may be the input to thousands of other neurons. Learning occurs by repeatedly activating certain neural connections over others, and this reinforces those connections. This makes them more likely to produce a desired outcome given a specified

input. This learning involves feedback when the desired outcome occurs, the neural connections causing that outcome becomes strengthened. [32].

In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs. The connections between artificial neurons are called 'edges'. Artificial neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer (the input layer) to the last layer (the output layer), possibly after traversing the layers multiple times.

The original goal of the ANN approach was to solve problems in the same way that a human brain would. However, over time, attention moved to performing specific tasks, leading to deviations from biology.

2.3.2. Elements of Artificial Neural Networks

A. Layers: Basically, there are 3 different layers in a neural network

- **Input Layer:** The Input layer communicates with the external environment that presents a pattern to the neural network. Its job is to deal with all the inputs only. The input layer provide information from the outside world to the network using input neurons which are inside it. Every input neuron should represent some independent variable that has an influence over the output of the neural network.
- **Hidden Layer:** The hidden layer is the collection of neurons which has activation function applied on it and it is an intermediate layer found between the input layer and the output layer. Its job is to process the inputs obtained by its previous layer. So it is the layer which is responsible extracting the required features from the input data. Hidden layer also has hidden neurons.

- **Output Layer:** The output layer of the neural network collects and transmits the information accordingly in way it has been designed to give. The pattern presented by the output layer can be directly traced back to the input layer. The number of neurons in output layer should be directly related to the type of work that the neural network was performing.

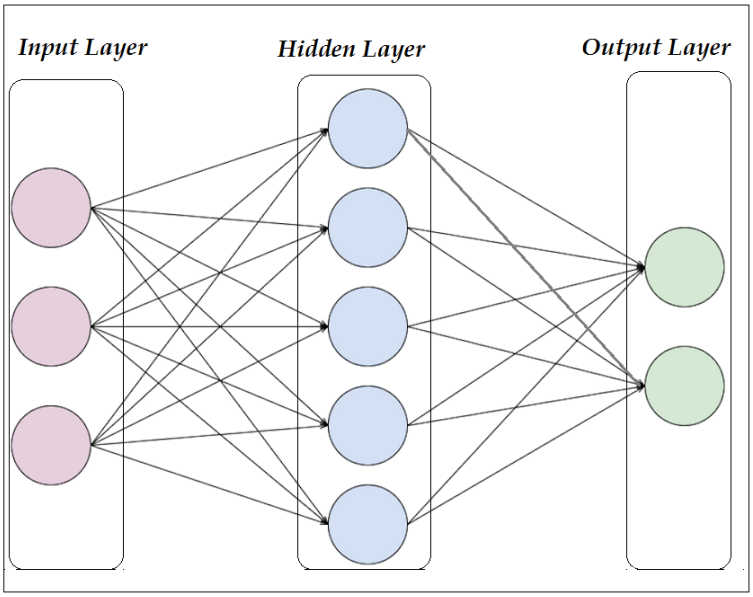


Figure 1. 5 Layers in Simple Artificial Neural Network

B. Neurons: also called nodes, is the basic building block of a NN. Input layer, hidden layer, and output layer have neurons.

C. Weight: are the most important factor in converting an input to impact the output. Each neuron is connected to every other neuron by means of directed links. Links are associated with weights. Weights contain information about the input signal and is represented as a matrix. Weights are numerical parameters which determine how strongly each of the neurons affects the other.

D. Bias: is an additional parameter which is used to adjust the output along with the weighted sum of the inputs to the neuron. Bias is like another weight. It's included by adding a component x0=1 to the input vector X.

X=(1, x1, x2…xi,…xn). Bias is of two types; positive bias increase the network input, negative bias decrease the network input.

E. **Activation functions**: also called transfer function. Each neuron has an activation function that defines the output of the neuron. We use the activation functions to propagate the output of a neuron forward. This output is received by the neurons of the next layer to which this neuron is connected. Used to calculate the output response of a neuron. Sum of the weighted input signal is applied with an activation to obtain the response. Among the most common activation function; Linear, Tan-Sigmoid, Tanh, Softmax, and ReLu are found.

F. **Learning Algorithm:** is an automatic adaptive method that tries to fit the appropriate weights to the system solution. A learning algorithm often determines the values of the weights [33]. Learning algorithms have different characteristics and performance in terms of memory requirements, speed, and precision. Below, the most common learning algorithms are described. *The following descriptions are according to [34]*.

   - **Gradient Descent:** also known as steepest descent, is the simplest training algorithm. It requires information from the gradient vector, and hence it is a first order method.
   - **Levenberg-Marquardt:** also known as the damped least-squares method, has been designed to work specifically with loss functions which take the form of a sum of squared errors. As we have seen the Levenberg-Marquardt algorithm is a method tailored for functions of the type sum-of-squared-error. That makes it to be very fast when training neural networks measured on that kind of errors. The Levenberg-Marquardt algorithm is recommended when we have small data sets.

G. **Learning Rate:** The learning rate is a hyper parameter that controls how much to change the model in response to the estimated error each time the model weights are updated. The learning rate controls how quickly the model is adapted to the problem. Smaller learning rates require more training epochs given the smaller changes made to the weights each update, whereas larger learning rates result in rapid changes and require fewer training epochs.

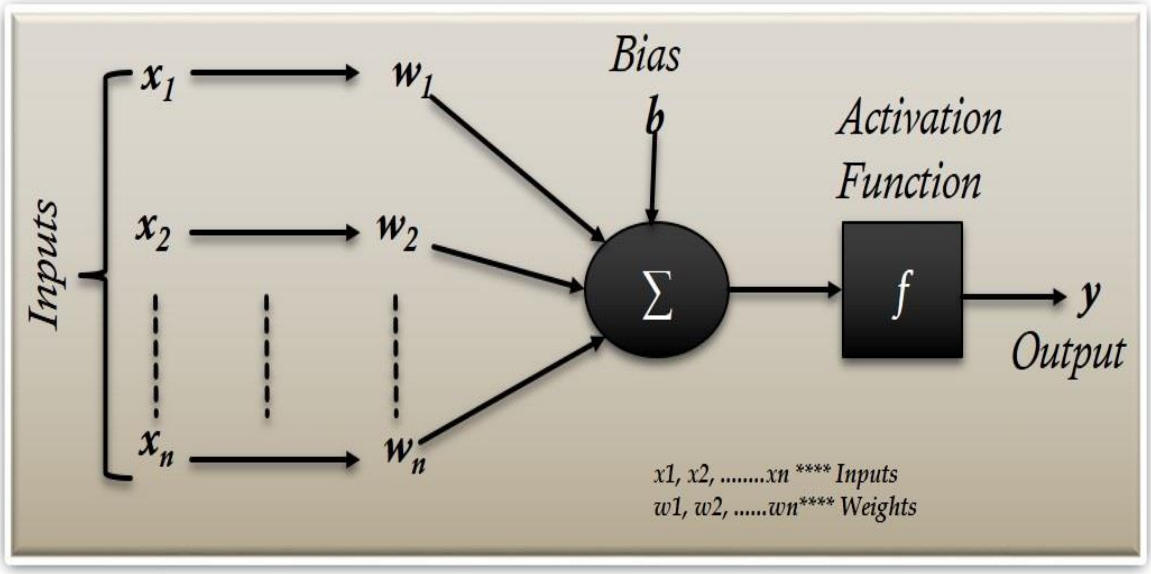


Figure 1. 6 Elements of Artificial Neural Network

2.3.3. ANN Architectures

Network architecture is the arrangement of neuron to form layers and connection patterns formed within and between layers. In ANN the artificial neurons are connected in many different ways forming architectural characteristics. The learning algorithms and the architectures are closely related [33]. It is important to have a clear concept of how the artificial neurons may be interconnected to form the specific architectures because this will define how the computer implementations of the architectures can be done. Next, explanation of the most common ANN architectures is as follows.

2.3.3.1. Single-Layer Feedforward Networks

In this simplest network, a layer of input neurons is connected to a layer of output neurons. The "single-layer" designation refers to the output layer. The layer of input neurons is not considered because it does not process any computation over the input values. It just bypasses the input values [33]. Output layer is formed when different weights are applied on input nodes and the cumulative effect per node is taken. After this, the neurons collectively give the output layer compute the output signals.

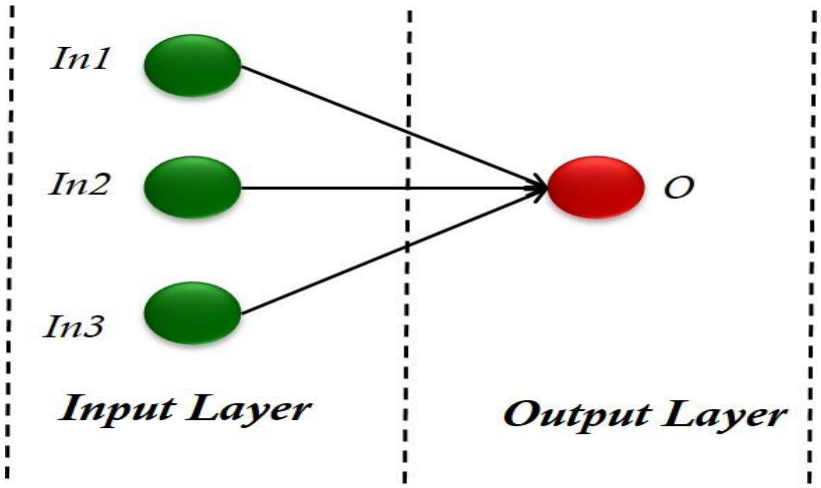


Figure 1. 7 Single-Layer Feedforward Network

2.3.3.2. Multi-Layer Feedforward Networks

This layer also has hidden layer which is internal to the network and has no direct contact with the external layer. Existence of one or more hidden layers enable the network to be computationally stronger. There are no feedback connections in which outputs of the model are fed back into itself.

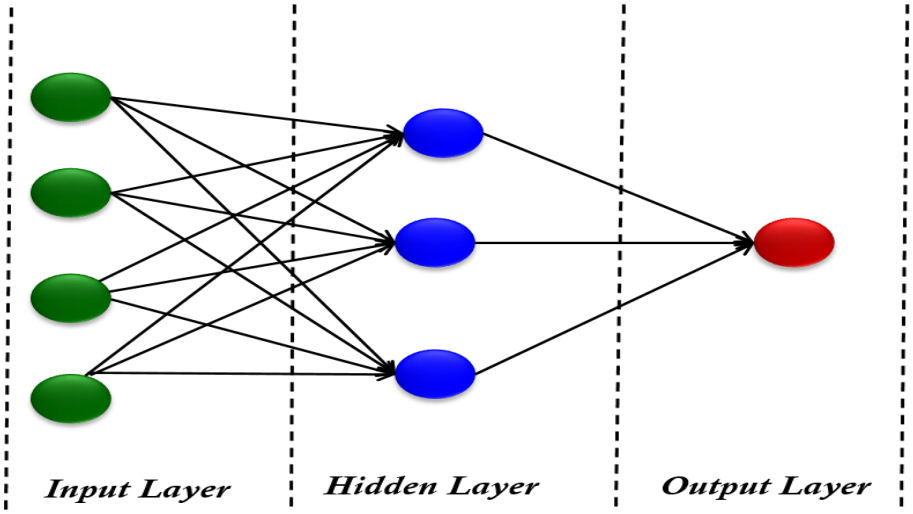


Figure 1. 8 Multi-Layer Feedforward Network

2.3.3.3. Recurrent Networks

This network model has at least one feedback loop. It may have the same architecture as a layered network, but it is necessary to have the feedback. The feedback can happen from the output of one neuron back to the input of another neuron [33]. Recurrent neural networks (RNNs) are a strong choice for building up sequential representations of data over time: tasks such as handwriting recognition and voice recognition.

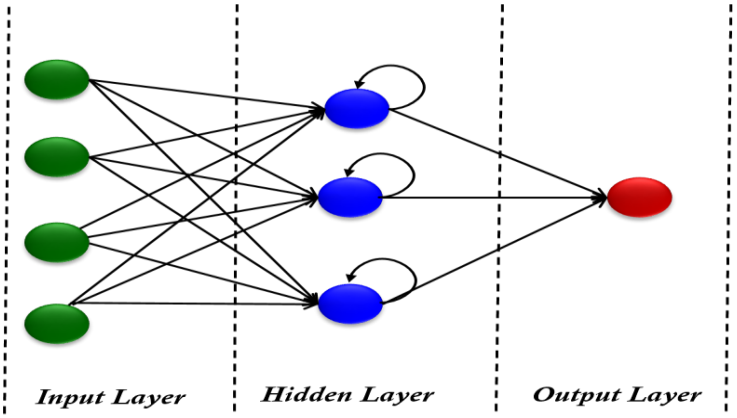


Figure 1. 9 Recurrent Neural Network Architecture

2.3.4. Application of Artificial Neural Network

One of the greatest advantages of artificial neural networks is their capability to learn from their environment. Learning from the environment comes useful in applications where complexity of the environment (data or task) make implementations of other type of solutions impractical. As such artificial neural networks can be used for variety of tasks like classification, function approximation, data processing, filtering, clustering, compression, robotics, regulations, decision making, etc. Choosing the right artificial neural network topology depends on the type of the application and data representation of a given problem. When choosing and using artificial neural networks we need to be familiar with theory of artificial neural network models and learning algorithms [35].

Followings are some of the areas, where ANN is being used. It suggests that ANN has an interdisciplinary approach in its development and applications [36].

Speech Recognition: Speech occupies a prominent role in human-human interaction. Therefore, it is natural for people to expect speech interfaces with computers. In the present era, for communication with machines, humans still need sophisticated languages which are difficult to learn and use. ANN is playing a major role in this area.

Character Recognition: It is an interesting problem which falls under the general area of Pattern Recognition. Many neural networks have been developed for automatic recognition of handwritten characters, either letters or digits.

Signature Verification Application: Signatures are one of the most useful ways to authorize and authenticate a person in legal transactions. Signature verification technique is a non-vision based technique. For this application, the first approach is to extract the feature or rather the geometrical feature set representing the signature. With these feature sets, we have to train the neural networks using an efficient neural network algorithm. This trained neural network will classify the signature as being genuine or forged under the verification stage.

Human Face Recognition: It is one of the biometric methods to identify the given face. It is a typical task because of the characterization of "non-face" images. However, if a neural network is well trained, then it can be divided into two classes namely images having faces and images that do not have faces. First, all the input images must be preprocessed. Then, the dimensionality of that image must be reduced. And, at last, it must be classified using neural network training algorithm. Following neural networks are used for training purposes with preprocessed image.

2.4. Amharic Language

Amharic is one of the Ethiopian Semitic languages which are a subgrouping within the Semitic branch of the Afro-Asiatic languages. The language serves as the official working language of Ethiopia, and the working language of several of the states within the Ethiopian federal system. According to the 1999 census, Amharic is spoken by 21.6 million native speakers and 4 million secondary speakers in the country. Amharic is the second most commonly spoken Semitic language in the world after Arabic. Additionally, 3 million

emigrants outside of Ethiopia speak the language. In Washington DC, Amharic become one of the six non-English languages in the language access act of 2004 G.C, which allows government services and education in Amharic [37]. Throughout the country, Amharic is taught as a subject in schools. Many universities have also departments of Amharic to train Amharic teachers, journalists, and people who work in public relations. Currently, different mass Media like radios, television broadcasts and the press in the country use Amharic for disseminating information to the public [38].

2.4.1. Amharic Phonetics

Phonetics is the study of speech sounds and their production, classification, and transcription. Phonology, on the other hand, is the study of the distribution and patterning of speech sounds in a language and of the tacit rules governing pronunciation. Articulatory phonetics deals with how the human vocal apparatus is manipulated to produce sounds. The basic assumption of articulatory phonetics is that sounds are best described in terms of the configurations of the vocal tract necessary to utter the sounds. Sounds can also be classified as vowels and consonants which are produced in different ways. While consonants are articulated with a substantial degree of obstruction in the oral cavity, vowels are produced with a relatively free airflow [38]. Articulatory phonetics shows that characteristics of sound are determined by the position of the various articulators in the vocal tract and the state of the vocal cords [39]. Three features are used to express the phonetics of Amharic language: Voicing, Manner, and Place of articulation. Linguists decompose a spoken language into elements of linguistically distinct sounds called phonemes. According to [39] Phonemes are determined and classified according to their corresponding articulatory configurations. Based on the voicing aspect sounds classified as voiced and unvoiced. In the articulation of manner, sounds are classified in the classes: Stops, Fricatives, and Affricates. Place of articulations includes Labial, Dental, Palatal, Velar, and Glottal. Sounds can also be classified as vowels and consonants. Vowel and consonant sounds are produced in fundamentally different ways.

2.4.1.1. Amharic Consonants

According to [39] Consonants, as opposed to vowels, are characterized by significant constriction or obstruction in the pharyngeal and/or oral cavities, that some consonants are voiced; others are not. A set of 34 phones, seven vowels, and 27 consonants and 7 other repeated consonants, makes up the complete inventory of sounds for the Amharic language. Amharic consonants according to the manner of articulation are generally classified as stops/ዝግ/, fricatives/ተሸለቀቂ/- formed when the stricture is very narrow (but without total closure) so that when air flows out, a hissing noise is made, and lateral- is made when air flows out of the sides of the mouth, nasals/ የአፍንጫ/ - air to flow out of the nose, liquids/ የምላስ/, and semi-vowels. Table 2.1 shows the phonetic representation of the consonants of Amharic as to their manner of articulation, voicing, and place of articulation are of voiceless //የማይነዝር/, voiced/ ነዛሪ/, and glottal/ የጉሮሮ/.

Amharic consonants are generally classified based on their voicing, manner, and place of articulation. In articulatory phonetics, the place of articulation (also point of articulation) of a consonant is the point of contact where an obstruction occurs in the vocal tract between an articulatory gesture, an active articulator (typically some part of the tongue), and a passive location (typically some part of the roof of the mouth). Table 2.1 shows the phonetic representation of the consonants of Amharic as to their manner of articulation, voicing, and place of articulation.

Manner of articulation	**Voicing**	**Place of Articulation**						
		Bilabial	**Labiodental**	**Alveolar**	**Palatal**	**Velar**	**Labiovelar**	**Glottal**
Stops	**Voice**	ብ [b]		ድ [d]		ግ [g]	ጓ [g^w]	
	Voiceless	ፕ [p]		ት [t]		ክ [k]	ኳ [k^w]	ዕ[?]
	Ejective	ጵ [p’]		ጥ [T]		ቅ [k’]	ቋ [k’^{w}]	
Fricatives	**Voice**		ቭ[v]	ዝ [z]	ዥ [z’]			
	Voiceless		ፍ[f]	ስ [s]	ሽ [S]			ህ [h] ኋ [h^w]
	Ejective			ጽ [s’]				
Affricates	**Voice**				ጅ [j]			
	Voiceless				ች [c]			
	Ejective				ጭ [c’]			
Nasals		ም [m]		ን [n]	ኝ [N]			
Liquids				ል [l]	ር [r]			
Glides		ው [w]			ይ [y]			

Table 1. 1 Amharic Consonants

Dictionary Definition	
Stops	A consonant produced by stopping the flow of air at some point and suddenly releasing it
Fricatives	A continuant consonant produced by breath moving against a narrowing of the vocal tract.
Affricatives	A composite speech sound consisting of a stop and a fricative articulated at the same point.
Nasals	A consonant produced through the nose with the mouth closed.
Liquids	A frictionless continuant that is not a nasal consonant.
Glides	A vowel like the sound that serves as a consonant.
Bilabials	A consonant whose articulation involves the movement of both lips.
Alveolar	A consonant articulated with the tip of the tongue near the gum ridge.
Palatals	A semivowel produced with the tongue near the palate.
Velars	A consonant produced with the back of the tongue touching or near the soft palate.
Glottal	A consonant produced by the glottis.
Voiced	Produced with the vibration of the vocal cords.
Voiceless	Produced without vibration of the vocal cords.
Ejective	a non-pulmonic consonant formed by squeezing air trapped between the glottis and an articulator further forward, and releasing it suddenly

Table 1. 2 Dictionary Definition of Terms

2.4.1.2. Amharic Vowels

Vowels are open sounds, made largely by shaping the vocal tract rather than by interfering with the flow of air stream [40]. Vowels are most usefully described in terms of the position of the tongue as they are articulated [40]. A vowel articulated with the body of the tongue relatively forward is classified as a front vowel; one made with the body of the tongue relatively high is a high vowel. Vowels produced with the body of the tongue neither high nor low are called mid vowels [40]. Vowels produced with the tongue body front are called front vowels while those made with the tongue body back are called back vowels. Vowels accompanied by lip rounding as in (ኡ and ኦ) are called rounded vowels while the other vowels are called unrounded vowels. In general, Amharic vowels (ኧ, ኡ, ኢ, ኣ, ኤ, እ and ኦ) are categorized in to rounded (ኡ and ኦ) and unrounded (ኧ, ኡ, ኢ, ኤ and እ) based on their manner of articulation. The other categorization is based on the height of the tongue and part of the tongue which depends on their production as indicated in Table 2.2 [41].

	Front/Unrounded	**Central/ Unrounded**	**Back/Rounded**
High	ኢ [i]	እ [I]	ኡ [u]
Mid	ኤ [e]	ኧ [E]	ኦ [o]
Low		ኣ [a]	

Table 1. 3 Amharic Vowels

2.5. Related Works

(S. A. Mahmood and L. E. George, 2007), investigate neural based speaker recognition system. LPC has been used as a feature extraction method. And back propagation neural network has been used for the purpose of speaker modeling and identification. Achieved 90% accuracy for 10 speakers [43].

(M. S. Sinith et al. , 2010), emphasis on text-Independent speaker identification system using Mel-Frequency Cepstral Coefficients (MFCC) as the speaker speech feature parameters in the system and the concept of Gaussian Mixture Modeling (GMM) for modeling the extracted speech feature. And used the Maximum Likelihood Ratio Detector algorithm for the decision making process. The experimental study has been conducted on MATLAB 7. Gaussian mixture speaker model attains high recognition rate for various speech durations. The recognition rate is maximum (98.8 %) when the speech is of 60 seconds duration and the number of Gaussians is 16 [44].

(M. Islam, F. Khan, and A. M. Haque, 2013), presents the implementation of Text Independent Speaker Identification system. Feature extraction task has been done using Mel Frequency Cepstral Coefficients (MFCC) acquisition algorithm that extracts features from the speech signal, which are actually the vectors of coefficients. The backpropagation algorithm of the artificial neural network stores the extracted features on a database and then identify speaker based on the information. Achieved near 100% accuracy in case of static speech signal and above 90% accuracy in case of real time speech signal [45].

(Amr Rashed, 2014), this paper proposed a fast algorithm for speaker recognition. This algorithm first records voice patterns of speakers via noisy channel and use some of noise removal techniques. The feature is extracted by Mel Frequency Cepstral Coefficient (MFCC) and the feature is reduced by Principal component analysis (PCA) technique. Then the result vector is fed to ANN classifier. Experimental results indicates that using ANN with weight/bias training algorithm have better performance. The result shows that the proposed algorithm achieved on average about 99% accuracy rate and higher speed rate in comparison with other methods [46].

(A. Antony and R. Gopikakumari, 2018), introduces an isolated word speaker identification system based on a new feature extractor and using ANN. The system is designed for both text independent and text dependent speaker identification system for English words. The speech is recorded using audio wave recorder. Then the preprocessing is applied for the given speech signals. UMRT is a transform which has been used for image compression. Combinations of MFCC and UMRT are taken and are used as a feature extractor. The classification of the features is done using Multi-layer perceptron with back propagation algorithm. The accuracy is taken using confusion matrix. The accuracy achieved is around 97.91% for speech dependent systems while for speech independent system the accuracy is around 94.44% [47].

(A. Azene, 2015), the first attempted Amharic language speaker recognition research. It presents text-independent speaker identification system for the Amharic language. Speech signals are collected from different speakers including both sexes as well as different age groups. MFCC had been used to extract features from the speech signals and to generate feature vector. VQ and GMMs had been used for training and identification purpose. The intention of the researcher was to see which modeling approach is better for text-independent speaker identification. For a total of 50 speakers, 74.2% accuracy was achieved when VQ approach is used where as 84.3% accuracy for the GMMs. The researcher tried to see the speaker identification accuracy based on gender. 25 male and 25 female speakers were considered. From the experiment, 86.2% and 85.9% accuracy was achieved for male and female speakers respectively [5].

(A. D. Mengistu, 2017), presents an automatic text-independent speaker identification system for the Amharic language in noisy environments. Speech signals are collected from different 100 speakers including both gender. Each speech has 10 seconds duration from each individual. Combination of MFCC, LPCC, and GFCC had been used for feature extraction purpose. VQ and GMMs had been used for training and identification purpose. The researcher was attempting to see which modeling approach is better for text-independent speaker identification with the combination of the three feature extraction techniques (MFCC, LPCC, and GFCC). The researcher conducted two experiments; **One**, VQ modeling approach with the combination of the three feature extraction techniques for

30, 60, and 90 speakers. And achieved 77.2%, 70.9%, and 69% accuracy respectively. **Two**, GMM modeling approach with the combination of the three feature extraction techniques for 30, 60, and 90 speakers. And achieved 75.2%, 76.9%, and 78% accuracy respectively [6].

(A. D. Mengistu D. Melesew, 2017), presents a hybrid approach of VQ and GMM have been used for classifying dialects of Amharic language. In speech signals collection, total of 100 speakers from each group of dialects (Gojjam, Wollo, Shewa, and Gonder) are considered. MFCC feature vectors are used to recognize the dialects of speakers. When 25 speakers are considered from areas, 85.9% accuracy had been achieved. When the number of speakers are increased to100, which is the maximum number of dialect speakers of the experiment, 92.7% accuracy had been achieved [7].

No	Author(s)	Title and Year	Speaker Recognition classification		Feature Extraction Techniques	Modelling Techniques & Toolkits	Findings
			Classification	**Category**			
1	M.S.Sinith et.al	*A Novel Method for Text-Independent Speaker Identification Using MFCC and GMM, **2010***	*Speaker Identification*	*Text-Independent*	MFCC	GMM, MATLAB	Recognition rate is maximum (98.8 %) when the speech is of 60 seconds duration and the number of Gaussians is 16.
2	Md. Monirul Islam et.al	*A Novel Approach for Text-Independent Speaker Identification Using ANN, **2013***	*Speaker Identification*	*Text-Independent*	MFCC	ANN, MATLAB	Near 100% accuracy in case of static speech signal and above 90% accuracy in case of real time speech signal.
3	Amr Rashed	*Fast algorithm for noisy speaker recognition using ANN, **2014***	Speaker *Identification*	*Both*	MFCC	ANN, MATLAB	Proposed algorithm achieved on average 99% accuracy
4	Aykefam Azene	*Text-independent speaker identification system for Amharic language, **2015***	*Speaker Identification*	*Text-Independent*	MFCC	VQ & GMM, MATLAB	74.2% & 84.3% accuracy had been achieved for 50 speakers using VQ & GMMs respectively.

5	Abrham Debasu	*Automatic Text Independent Amharic Language Speaker Recognition in Noisy Env't Using Hybrid Approaches of LPCC, MFCC and GFCC,* ***2017***	*Speaker Identification*	*Text-Independent*	LPCC, MFCC, and GFCC	VQ & GMM, MATLAB	77.2%, 70.9%, and 69% accuracy had been achieved for 30, 60, 90 speakers respectively using VQ and 75.2%, 76.9% and 78% accuracy had been achieved for 30, 60, 90 speakers respectively using GMMs.
6	Abrham D. and Dagnachew M.	*Text Independent Amharic Language Dialect Recognition: A Hybrid Approach of VQ and GMM,* ***2017***	*Speaker Identification*	*Text-Independent*	MFCC	VQ & GMM, MATLAB	85.9% and 92.7% accuracy had been achieved for 25 & 100 speakers respectively.
7	Anett Antony and R. Gopikakumari	*Speaker identification based on combination of MFCC and UMRT based features,* ***2018***	*Speaker Identification*	*Both*	*MFCC and UMRT*	ANN, MATLAB	97.91% and 94.44% accuracy had been achieved for speech dependent and speech independent systems respectively

Table 1. 4 Summary of Related Works

CHAPTER THREE: METHODODOLGY

3.1. Framework of Methodology

Activities involved in this study framework of methodology are speech collection, preprocessing of speech signals, feature extraction of preprocessed data, feature modeling of extracted features, feature matching during identification, and performance evaluation. All activities has been done using an appropriate techniques and tools. Figure 1.11 shows the overall detailed framework of the methodology.

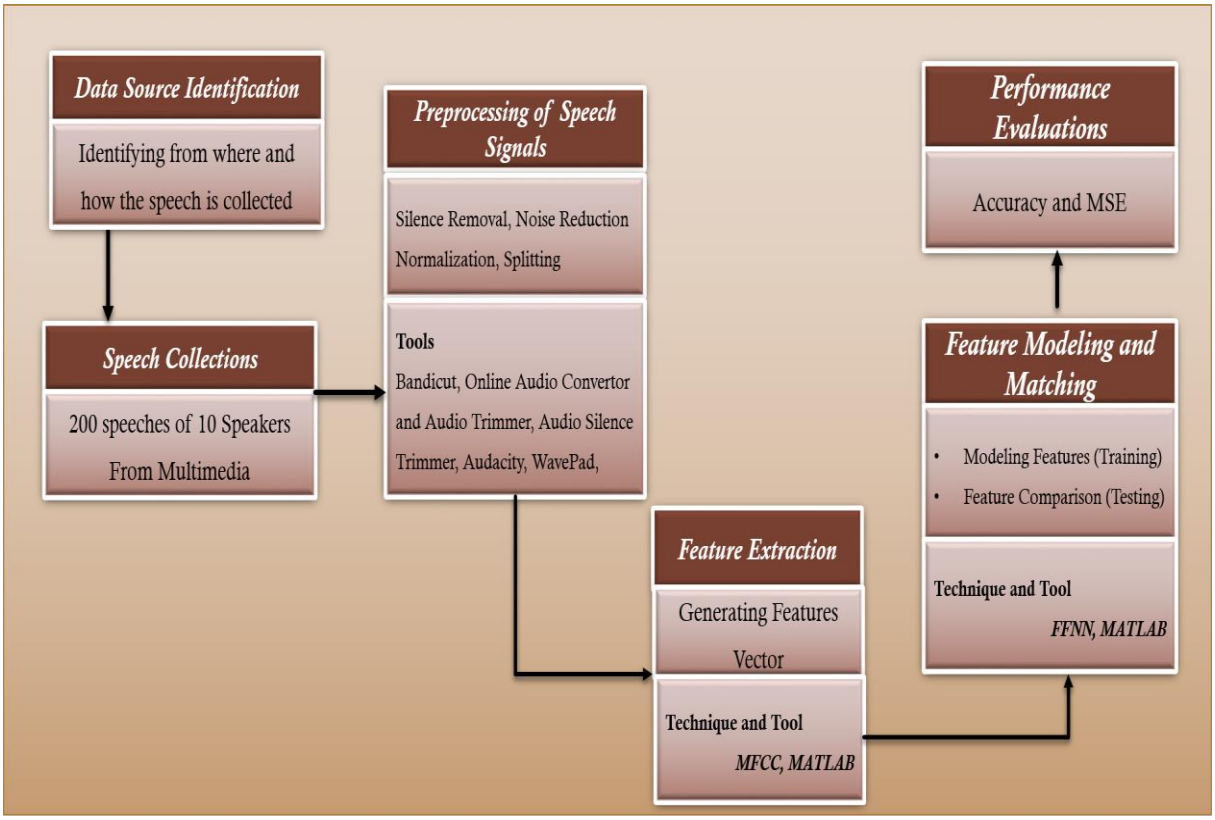


Figure 1. 10 Conceptual Framework of the Methodology

3.2. Research Design

Research design is defined as a framework of methods and techniques chosen by a researcher to combine various components of research in a reasonably logical manner so that the research problem is efficiently handled. Selecting an appropriate research design has a great impact for the research to be successful. The decisions make at this stage of the research is very essential to determine the quality of the conclusions can draw from research results. Because the expected outcome from this study is performance in percentage, the overall process was numerical. And since the internal key performance indicator selected for this study is *number of hidden neuron,* three experiments have been conducted. Therefore, the researcher has chosen quantitative and experimental research design to address the research questions mentioned in chapter one section 1.4.

3.3. Data Source Identification

As the study is speaker recognition and the required data is speech, the researcher was thought speeches from some volunteers who permit to be recorded. But, the researcher couldn't get adequate volunteers to record the speeches. So, multimedia was the speech data source in this study.

3.4. Speech Collections

A speech corpus is one of the fundamental requirements for any speech and speaker recognition research. The speech corpus is a collection of speech recordings which is accessible in computer readable form, and which has annotation and documentation sufficient to allow re-use of data. Apart from the specifics of the language itself, the main problem with doing speaker recognition for an under-resourced language like Amharic is the lack of previously available data. Thus, Ten Ethiopian famous people public speeches has been taken from multimedia (youtube). All of the speakers are native showa Amharic accent speakers. Public speech of each speaker is downloaded from youtube and cropped 30 seconds of speech from each speaker's speech using *Bandicut* software. Then, converted each 30 seconds video speech in to **44.1.KHz** sampling rate 32 bit float bit depth ***(.wav)*** format audio, using online audio convertor. Totally, 30 seconds duration speech of each

speaker has been prepared for training and testing separately from the same public speech based on time difference.

3.5. Preprocessing

Pre-processing is the first step for signal processing technique of speech. The preprocessing in speaker recognition discussed in this study includes silence removal, noise reduction, normalization, and speech signal splitting.

Figure 1. 11 AA.wav Original Audio

3.5.1. Silence Removal

Silence removal is the process of removing silence or unvoiced from the speech signal. While speaking, different people take a gap of different duration between words. In speaker recognition process, the essential thing is a unique feature to discriminate one speaker from other. Since silence is the same feature for all speakers, it is needless for recognition. So, silence should be removed from the speech signals. Because knowing the proper location of speech and silence or unvoiced helps not only reduces the amount of processing but also increases the accuracy of speech processing system. ***Audio silence trimmer software*** is used to remove silence at beginning, middle, and end of speech signal. The software is selected because it is free and easy to use. Here, the duration is decreased because the silence is removed from it. As per the duration of silence length, each speech signal wav file minimized in duration in the range of 21-26 seconds. To make them at equal duration,

20 seconds speech has been trimmed from all speech signals of each speaker using online audio trimmer.

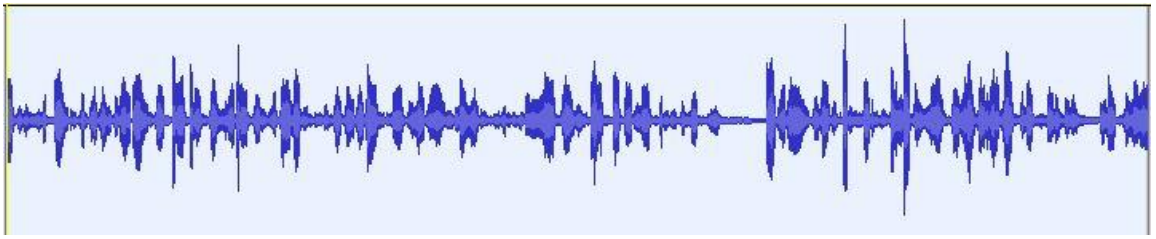


Figure 1. 12 AA.wav after Silence Removal

3.5.2. Noise Reduction and Normalization

Since the speech signals are taken from different environment and they are public speech, there is background noise with all speeches. These background noise should be removed in order to extract the exact unique features of the speakers'. So, after silence in the speech signals has been removed, the next step is noise reduction and normalization. Normalization is the application of a constant amount of gain to an audio recording to bring the amplitude to a target level (the norm). Because during noise reduction, the loudness of the speech signal is decreased. Normalization changes back the loudness to normal (as it was before noise reduction). ***Audacity software*** is used to reduce background noise of the speech signal and to normalize. The reason behind selecting audacity is because it is one of top 10 (3rd) audio editing software [48], and open source.

Figure 1. 13 AA.wav after Noise Reduction

Figure 1. 14 AA.wav after Normalization

3.5.3. Splitting of Audio Speech Signals.

After silence and noise is removed from the speech signals, the 20 seconds wav audio is split in to 20 pieces. These is done for each speaker's speech. Totally, 200 speeches (*10 speakers * 20 pieces of speech*) is prepared for training and testing separately. Maximum duration of one speech is 1 second. Open source *WavePad Sound Editor software* is used to split the speech signals into pieces. Table 1.6 shows the final training and testing datasets after preprocessing.

No.	**Speaker's Name**	**Speaker Code**	**Training Dataset**	**Testing Dataset**
1	Dr. Abiy Ahmed	AA	AA_00.wav to AA_19.wav	tAA_00.wav to tAA_19.wav
2	Biniam Belete	BB	BB_00.wav to BB_19.wav	tBB_00.wav to tBB_19.wav
3	Birtukan Mideksa	BM	BM_00.wav to BM_19.wav	tBM_00.wav to tBM_19.wav
4	Fikadu T/Mariam	FT	FT_00.wav to FT_19.wav	tFT_00.wav to tFT_19.wav
5	Meaza Biru	MB	MB_00.wav to MB_19.wav	tMB_00.wav to tMB_19.wav
6	Dr. Mihret Debebe	MD	MD_00.wav to MD_19.wav	tMD_00.wav to tMD_19.wav
7	Muferiat Kamil	MK	MK_00.wav to MK_19.wav)	tMK_00.wav to tMK_19.wav)
8	Sahlework Zewde	SZ	SZ_00.wav to SZ_19.wav	tSZ_00.wav to tSZ_19.wav
9	Ustaz Abubeker Ahmed	UAA	UAA_00.wav to UAA_19.wav	tUAA_00.wav to tUAA_19.wav
10	Yetnebersh Nigussie	YN	YN_00.wav to YN_19.wav	tYN_00.wav to tYN_19.wav

Table 1. 5 Training and Testing Datasets

3.6. Feature Extraction

Speech is a highly complex signal, containing multiple streams of information. The aim of this transformation is to obtain a new representation which is more compact, less redundant, and more suitable for statistical modeling and the calculation of a distance or any other kind of score [14]. Simply, Feature extraction is the process of changing the speech signal to a set of feature vectors. It is the most important part of speaker recognition since it plays an important role in separate one speech from others. In this study, MFCC is used as feature extraction techniques to extracting features from speech signal. MFCC features are chosen based on their ability to extract discriminant information from speech, and the most commonly used acoustic features extraction technique for speech and speaker recognition.

- **Mel-Frequency Cepstrum Coefficients (MFCC)**

It is most preferable technique because it estimates the human hearing system response more nearly than other system. It takes human perception sensitivity with respect to frequencies into consideration [25]. *The* core idea is to split the signal into frames and apply a hamming window for each frame. Then, a spectral feature vector is generated for each frame by applying the FFT for each frame. Finally, the log of amplitude spectrum must be maintained and the spectrum is smoothed using a discrete cosine transform (DCT) to obtain cepstral features for each frame. The final goal of this stage is generating cepstral ***feature vector*** of the speech signal. Figure 3.5 shows steps in MFCC feature extraction.

Step 1: Framing

Since the speech framing duration standard is 20-40ms, the framing duration used in this study is 30ms. So, the frame length for a 44.1 kHz signal is ***0.03 * 44100 = 1323*** samples. Also, a shorter frame shift is chosen to track the continuity in the speech signal and not miss out any unforeseen changes at the edges of the frames. Quarter of the frame length has been taken for frame shift i.e., (***floor (1323/4) =330 samples***), which allows some overlap to the frames. The first 1323 sample frame starts at sample 0, the next 1323 sample frame starts at sample 330, etc. until the end of the speech file is reached. Therefore, **N=1323, M=330**

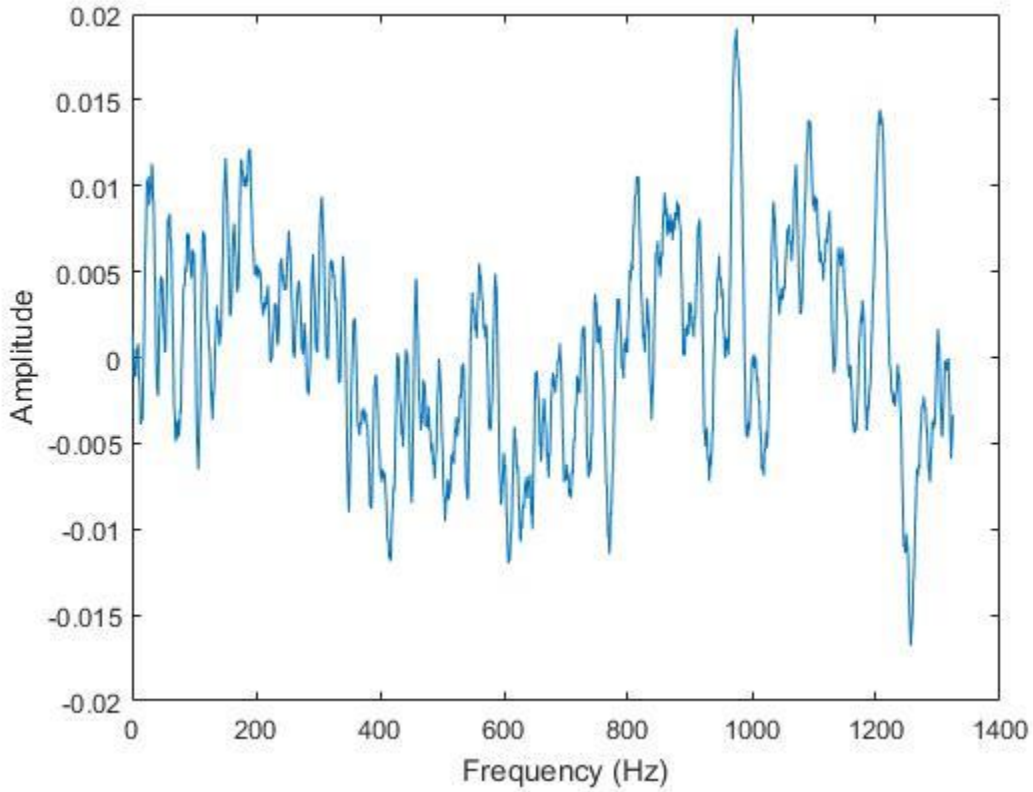


Figure 1. 15 Single Frame of AA_00.wav

Step 2: Windowing

As we have mentioned above, windowing technique is used to minimize the signal discontinuity. Each of the above frames is multiplied with a window function in order to keep continuity of the signal. So to reduce this discontinuity in this study, hamming window function is applied.

Hamming window = frame * hamming (length (frame));

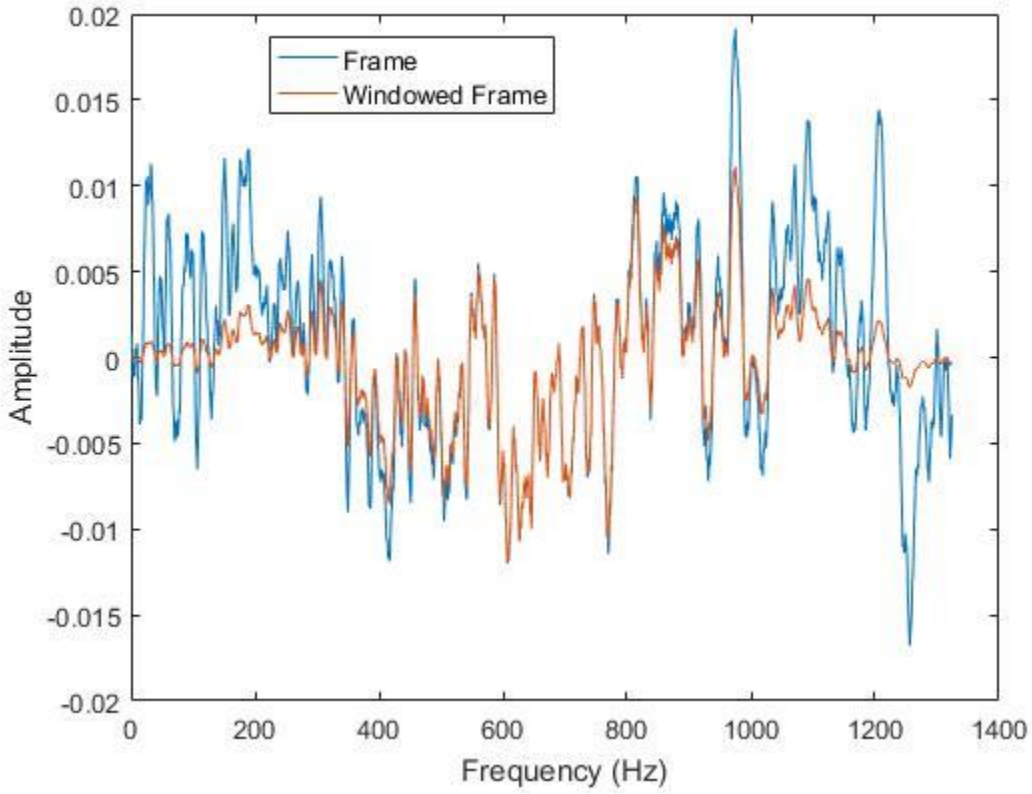


Figure 1. 16 Frame and Windowed Frame of AA_00.wav

Step3:

Step 3: Fast Fourier Transform

Fast Fourier Transform is used to convert each frame of N samples from time domain to frequency domain. FFT is perform to obtain magnitude of frequency response of each frame.

```
FFT= (abs (fft(frame)))
```

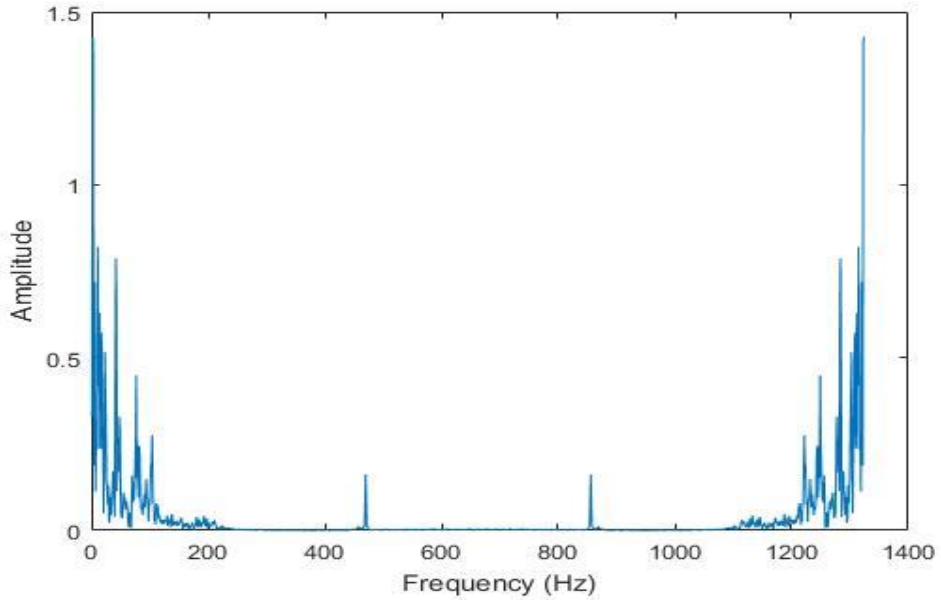


Figure 1. 17 FFT of single Windowed Frame of AA_00.wav

Step4: Mel Filter Bank Processing

The range of frequencies is very wide in FFT and voice signal does not follow the linear scale. The output of FFT is multiplied by a set of 32 triangular band-pass filter to get log energy of each triangular band-pass filter. Triangular band pass filters are used to extract the spectral envelope, which is constituted by dominant frequency components in the speech signal. The below equation is used to convert linear scale frequency into Mel scale frequency.

$$F(Mel) = 2595 * \log 10 (1+f/ 700)$$

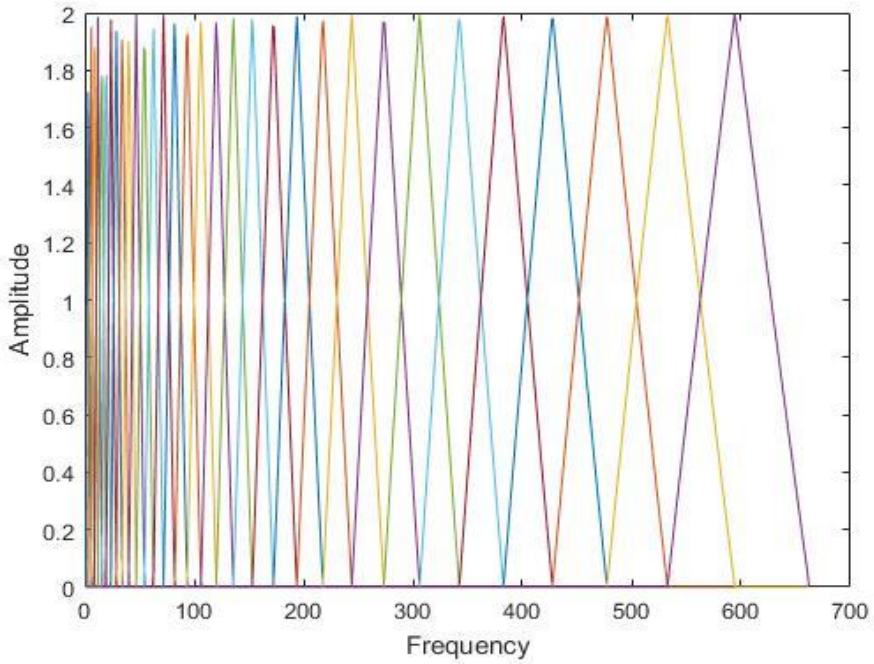


Figure 1. 18 Mel-Filter Bank of Single Windowed Frame of AA_00.wav

Step 5: Discrete Cosine Transform

Discrete cosine transform is the process to convert the log Mel spectrum into time domain. The result is called Mel Frequency Cepstrum Coefficients.

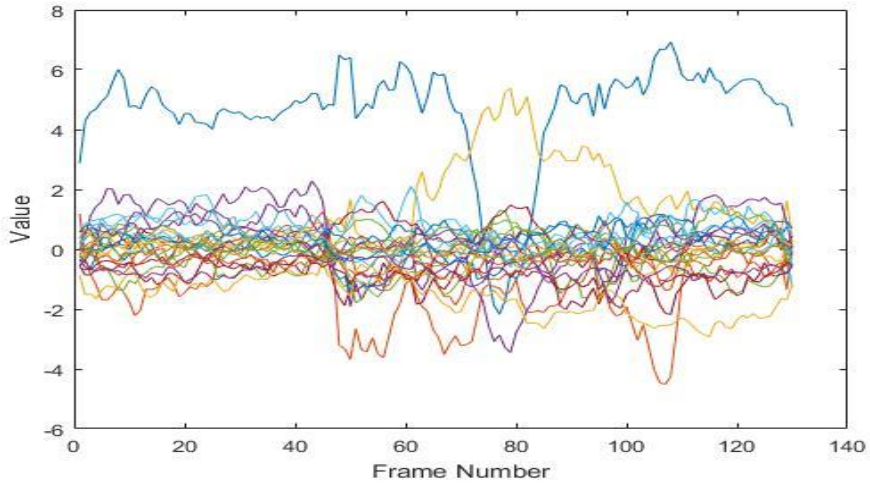


Figure 1. 19 MFCC Feature of AA_00.wav

All of the speech signals has been converted in to their respective feature vector by following the above five MFCC feature extraction technique steps. The number of feature coefficient to be extracted is 26. Finally, the feature vectors of the speech signals has been stored in to database.

3.7. Feature Modeling and Matching Technique

Once, after speaker feature vectors are stored in to database based on these features extraction; a sound represented as a sequence of feature vectors in the sound database. Each speakers' feature vector has trained for modeling technique. For modeling the feature vectors, neural network were used. Many architectures are possible for the neural networks. The architecture selected for this system was the **multi-layer feed-forward neural network.**

The feedforward net network is a specialized version of the multilayer feed-forward neural network architecture. The network was trained using the **trainlm** function. Trainlm is a network training function that updates weight and bias values according to ***Levenberg-Marquardt*** learning algorithm. The training algorithm start by initializing weights, biased, then summing the inputs multiply weights then after input training matrix is applied to the network pass it to activation function to extract the scaled output value. This output is then compared to the target matrix and assigns values to the different classes. The activation function use in this study was ***tansig***. As the training algorithm continues, the weights in the network are adjusted in order to minimize any errors in recognition. The training iterations continue, if not satisfied with the goal till to reaching best accuracy.

- **Input Layer**

The inputs for the neural network are matrices **26** feature vector so, input layer has 26 neurons. The 200 speech signals' feature vector divided in to 70% for training, 15% for validation and 15% for testing sets using **dividerand** function which divided randomly.

- **Hidden Layer**

Deciding the number of neurons in hidden layer is up to the researcher. Deciding the number of neurons in the hidden layer is a very important part of deciding your overall neural network architecture [49]. Even though hidden layers do not directly interact with the external environment but still they have a great influence on the final output. There are various approaches to find number of hidden neurons in hidden layer. The researcher used rule of thumb approach.

Rule of thumb [49]**.**
Rule of thumb method is for determining the correct number of neurons to use in the hidden layers, such as the following:
Rule 1: The number of hidden neurons should be in the range between the size of the input layer and the size of the output layer
Rule 2: The number of hidden neurons should be 2/3 of the input layer size, plus the size of the output layer
Rule 3: The number of hidden neurons should be less than twice the input layer size

Since there is three experiments based on number of hidden neuron, the researcher decided to select 15, 20, and 25 hidden neurons using the rule of thumb method.

- **Output Layer**

The number of neurons in output layer is depends on the number of classes stored in the database which is **10** neurons.

In this study, three experiments have been conducted. The experiments have been prepared from number of hidden neuron perspective.

- ✓ **Experiment 1:** Using 15 hidden neurons.
- ✓ **Experiment 2:** Using 20 hidden neurons.
- ✓ **Experiment 3:** Using 25 hidden neurons.

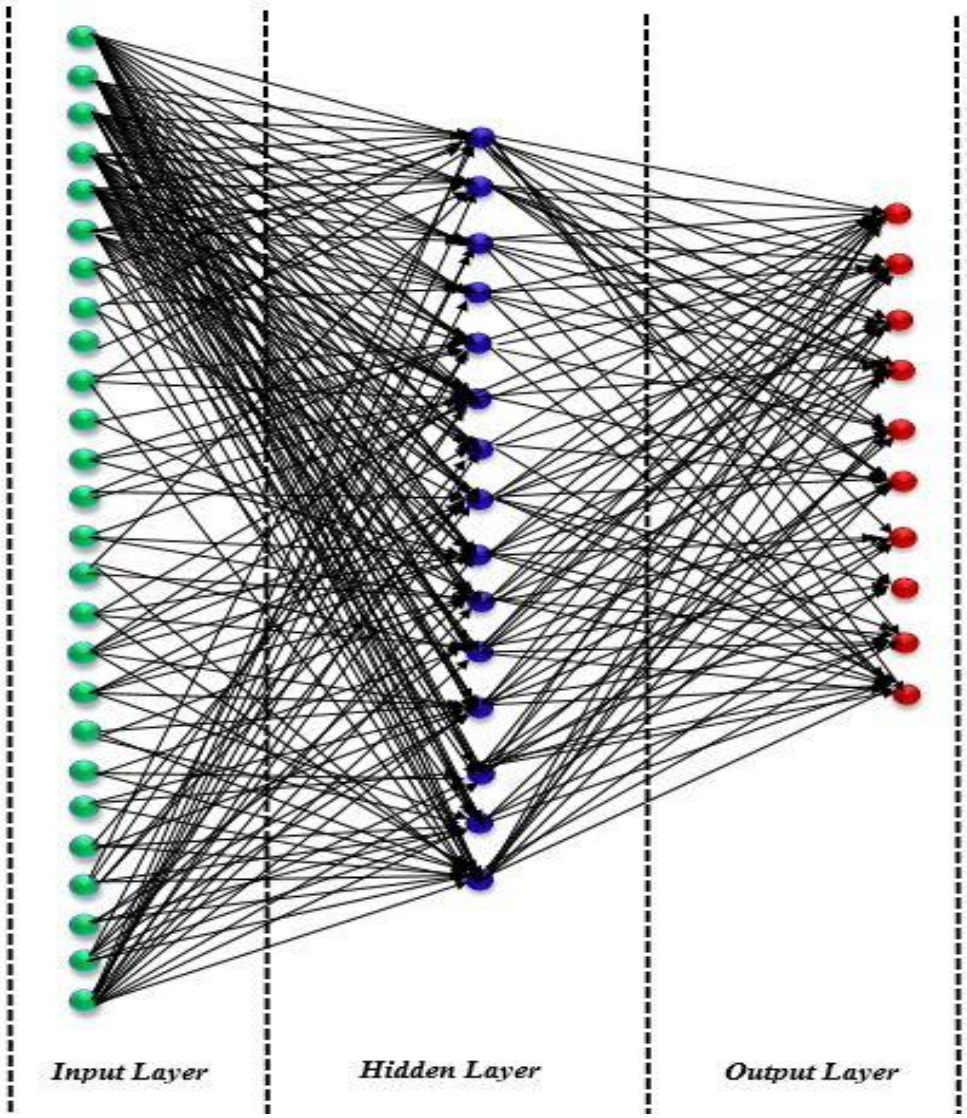


Figure 1. 20 Neural Network of Experiment One

The above neural network is to view the connection between 26 input neurons, 15 hidden neurons, and 10 output neurons. All the 26 input neurons supposed to be connected to 15 hidden neurons and the 15 hidden neurons supposed to be connected to 10 output neurons. However, connecting all the links to all neurons in the figure may lead us for connection invisibility. So, few links has been linked to the respective neurons to give a clue how the neural network has done.

3.8. Tools

The programming language and environment used to develop speaker recognition prototype based on neural networks is MATLAB version R2016a. MATLAB is a fourth-generation high-level programming language and interactive environment for numerical computation, visualization, and programming.

MATLAB is rich with different built-in toolboxes such as Neural Network and Signal Processing Toolboxes. Signal Processing Toolbox provides functions and apps to produce, measure, convert, filter, and visualize signals. Neural Network Toolbox provides tools for designing, implementing, visualizing, and simulating neural networks. It supports feedforward neural networks.

The code for the prototype was written in MATLAB script m-files. The speakers' sound database has been stored in the format of MAT with an extension .dat file format. Training and testing for the neural network were done in MATLAB with the association of the provided GUI.

The main reason to prefer MATLAB is having two basic features for the study experiment such as signal processing and neural network toolboxes. In addition to this, in general, it has a vast library, interactive environment, and built-in graphics for visualizing data and tools for designing with custom GUI.

3.9. Evaluation Methods

The recognition system must be evaluated or tested for its accuracy and error. In this study, performance evaluation is performed in two way. **1).** Based on the performance and confusion matrix plot generated during training of neural network. MATLAB Neural network toolbox provides a means of testing neural network model. **2).** Real time testing of speakers' speech that has been prepared for testing in to the trained neural network. Here, speakers' speech that is prepared for testing has been tested with respect to the trained neural network. The percentage of accurately recognized objects per each speaker has been computed.

3.10. Summary

No.	Techniques and Tools	Purpose
1	**Bandicut**	To cropped 30sec of speech from each speaker's speech
2	**www.online-audio-convertor.com**	To convert each 30 seconds video speech in to wav file
3	**www.audiotrimmer.com**	To trim 30 seconds wav file to 20 seconds.
4	**Audacity**	To reduce background noise of the speech signal and to normalize.
5	**Audio silence trimmer**	To remove silence at beginning, middle, and end of speech signal.
6	**WavePad**	To split the speech signals into pieces.
7	**MFCC**	To extract features from speech signals in order to generate feature vectors.
8	**FFNN**	To model the speaker, match features, and made identification decision.
9	**MATLAB R2016a**	A programming language and IDE on which all of the experiments have been done.

Table 1. 6 Summary of Techniques & Tools and their Purpose

CHAPTER FOUR: EXPERIMENTS, FINDINGS AND DISCUSSION

4.1. Experiments and Findings

4.1.1. Proposed Speaker Recognition Prototype

In order to have simple interaction in demonstration, the researcher developed graphical user interface prototype. So, in this study, the overall processes are conducted using GUI speaker recognition prototype (***see appendix C***).

Figure 1.21 depicts the general architecture of the proposed speaker recognition prototype which shows the overall process flow.

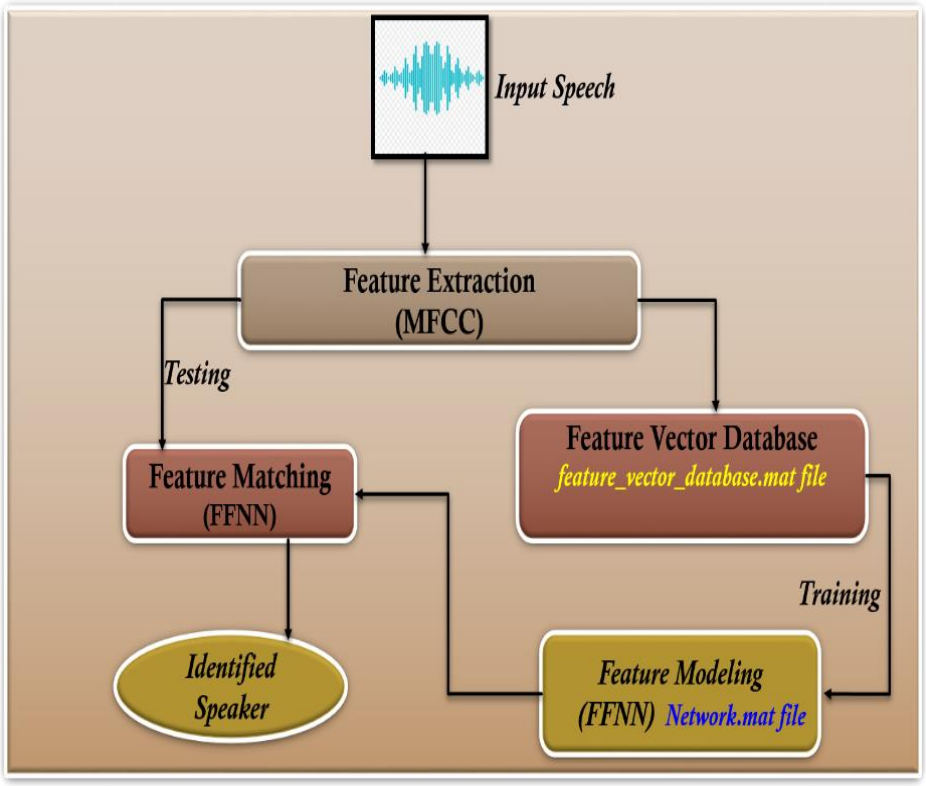


Figure 1. 21 Architecture of Proposed Speaker Recognition Prototype

Input speech is preprocessed speech signal. The preprocessed input speech always pass through feature extraction technique (MFCC). After feature extraction, there are two paths; training and testing. During training, feature vector of speech signal is created and stored in to database. Then, all feature vectors of speech signals which have been stored in the database will be trained for the feedforward neural network and ***'Network.mat'*** file which contain the trained neural network is created. During testing, a comparison between feature vector of the input speech and feature vector of speech signals in the trained neural network (in ***Network.mat***) will be done. Finally, after comparison has been made, the neural network make a decision to identify the speaker.

4.1.2. Feature Extraction

Since the prototype is GUI, the preprocessed input speech is loaded one by one to the system for feature extraction. The class ID is given when the input speech is loaded for feature extraction. 200 speech samples of 10 speaker has been loaded to the system for feature extraction and stored in database (***speech_feature_vector_database***).

features_data{14, 1}

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	2.8699	1.1832	-0.1893	-0.0076	-0.5339	-0.4064	-0.5626	-0.6033	-0.2004	-0.8687	-0.5650	-0.1731	0.3479
2	4.2790	-0.2889	0.0301	0.4275	-0.3201	-0.3705	-0.8595	-0.6920	0.1130	-1.5117	-0.3485	-0.3310	0.8741
3	4.5847	-0.8375	-0.0405	1.1251	-0.0074	-0.1313	-0.8534	-0.6244	0.2521	-1.4954	-0.0995	0.0944	0.9358
4	4.6909	-1.2093	-0.0015	1.4364	-0.3671	0.6279	-0.6668	-0.2150	0.1471	-1.6593	0.2909	0.4088	0.9851
5	4.9409	-1.6793	-0.2533	1.5911	-0.6909	1.0484	-0.5419	0.3667	0.4952	-1.3673	0.8776	0.2083	0.9911
6	5.1179	-1.6914	-0.1011	2.0299	-0.6555	0.7897	-0.6613	0.4115	0.6063	-1.2128	0.6449	0.1331	1.0622
7	5.5996	-1.5133	-0.3935	1.9989	-0.9916	0.7663	-0.6968	0.0636	0.6775	-1.2701	0.4354	0.5996	1.1428
8	6.0111	-1.2313	-0.5928	1.4986	-1.3707	0.9227	-0.5487	-0.1562	0.5108	-1.2437	0.6453	0.8413	1.1721
9	5.7036	-1.4455	-0.2355	1.8168	-1.3632	1.0735	-0.6215	-0.3802	0.2836	-0.9968	0.6953	0.8106	1.0788
10	4.7554	-1.7645	0.1953	1.8278	-1.0372	1.1053	-0.4247	-0.1545	0.4770	-0.7717	0.9953	0.2042	0.7368
11	4.8013	-2.2152	-0.0225	1.5327	-1.1544	1.0819	-0.6811	-0.1482	0.2963	-0.7788	1.2454	0.2078	0.9213
12	4.6836	-2.0491	0.3261	1.4392	-0.8911	1.2143	-0.9335	-0.0871	0.0290	-1.2212	1.2117	0.2859	1.1183
13	5.1574	-1.3266	0.3223	1.0983	-0.7902	1.2110	-1.1000	0.1735	0.1053	-1.5509	0.9598	0.0539	1.1092
14	5.4253	-1.3542	0.1687	1.4409	-0.6338	1.0628	-0.9483	0.1725	-0.1955	-1.6081	0.6595	-0.0656	1.0060
15	5.3100	-1.2870	0.3225	1.6961	-0.5680	0.8675	-1.1794	-0.2543	-0.0200	-1.4683	0.6605	-0.1907	1.0504
16	4.8300	-1.0823	0.6564	1.6126	-0.6194	1.1056	-1.0781	-0.3998	0.0179	-1.0867	0.9029	-0.3970	0.9124
17	4.6591	-0.9536	0.6255	1.4791	-0.3694	1.3657	-0.8097	-0.3308	-0.3412	-1.2785	1.0447	-0.3298	0.9709
18	4.5483	-1.1791	0.5441	1.5260	-0.2804	1.3737	-0.8132	-0.1179	-0.2073	-1.4184	1.2200	-0.2572	0.9381

Figure 1. 22 MFCC Feature Vector Matrix of AA_00.wav

Finally, the class ID is given based on the alphabetical order from 1 up to 10.

No.	Speaker Name	Speaker Code	Wav files' Name (20 per Speaker)	Class ID
1	Dr. Abiy Ahmed	AA	AA_00 to AA_19	**1**
2	Biniam Belete	BB	BB_00 to BB_19	**2**
3	Birtukan Mideksa	BM	BM_00 to BM_19	**3**
4	Fikadu T/mariam	FT	FT_00 to FT_19	**4**
5	Meaza Biru	MB	MB_00 to MB_19	**5**
6	Dr. Mihret Debebe	MD	MD_00 to MD_19	**6**
7	Muferiat Kamil	MK	MK_00 to MK_19	**7**
8	Sahilework Zewde	SZ	SZ_00 to SZ_19	**8**
9	Ustaz Abubeker Ahmed	UAA	UAA_00 to UAA_19	**9**
10	Yetnebersh Nigussie	YN	YN_00 to YN_19	**10**

Table 1. 7 Training dataset and their corresponding class ID after feature extraction

The code given for the speakers' is assigned using the two capital letters of first and last name of the speaker. Each speaker has 20 speeches, and the wav files are represented using the speaker code with numbers from **00-19** (***i.e. AA_00.wav, AA_01.wav, AA_02.wav……. AA_19.wav***).

4.1.3. Training of Neural Network

During the feature extraction stage, the speech signals are transformed in to feature vector in the way it will be suitable for training the neural network. The next step is training and testing the neural network. As tried to mention in chapter 3, in this study, three experiments have been conducted.

After setting the training parameters, the set of inputs with respective target outputs fed for the neural network sequentially. MATLAB has GUI neural networks tool. This GUI provides some useful plots to observe the Neural Network Training see below figure 1.23. In the GUI showed the time of training, a number of (trial) epochs, performance, gradient,

and validation checkers, and also provides the plots of the performance, training state, Regression, and confusion matrix.

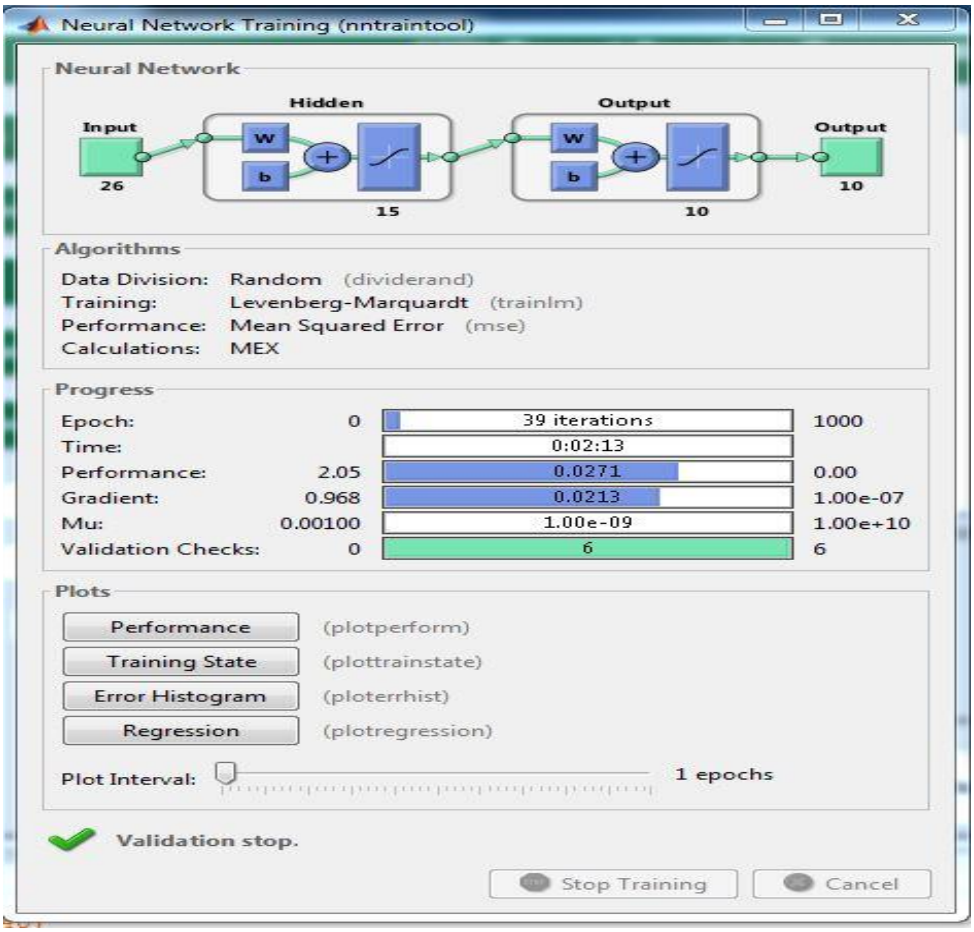


Figure 1. 23 Neural Network Training Tool; Training of Experiment One

The training parameters used for all experiments are listed below

- Number of Input Neurons=26
- Number of Hidden Neurons=15, 20, 25 (Expt.1, Expt.2, and Exp. 3 respectively).
- Number of Output Neurons=10
- Train Function= trainlm (Levenberg-Marquardt)
- Performance Function= Mean Squared Error (MSE)
- Divide Function= dividerand
- Training Ratio=70%, Validation Ratio=15%, Testing Ratio=15%
- Epochs=100

Figure 1.24 is the plot of the performance of experiment one for the training, validation, and testing versus the epochs. The blue, green, and red curve is for the training, validation, and testing performance respectively. The dotted black line represents the final goal and the dotted green line represents the best validation performance as a moving line to indicate best validation performance. Best validation performance met was **0.037317 at epoch 33.**

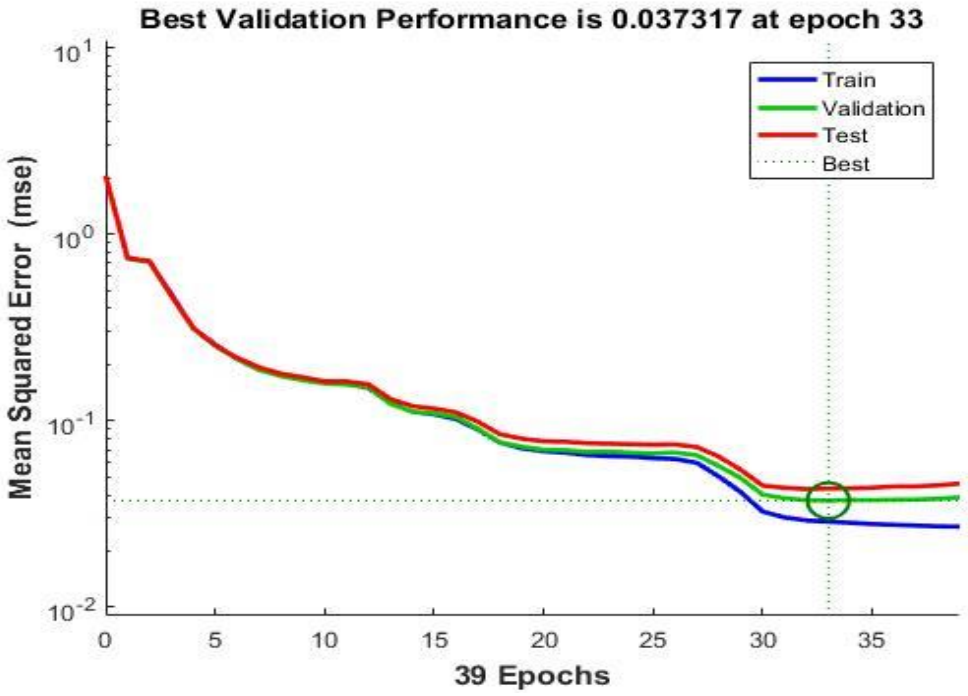


Figure 1. 24 Performance Plot of Experiment One [Hidden Neurons=15]

Figure 1.25 is the plot of the performance of experiment two for the training, validation, and testing versus the epochs. The blue, green, and red curve is for the training, validation, and testing performance respectively. The dotted black line represents the final goal and the dotted green line represents the best validation performance as a moving line to indicate best validation performance. Best validation performance met was **0.032121 at epoch 19.**

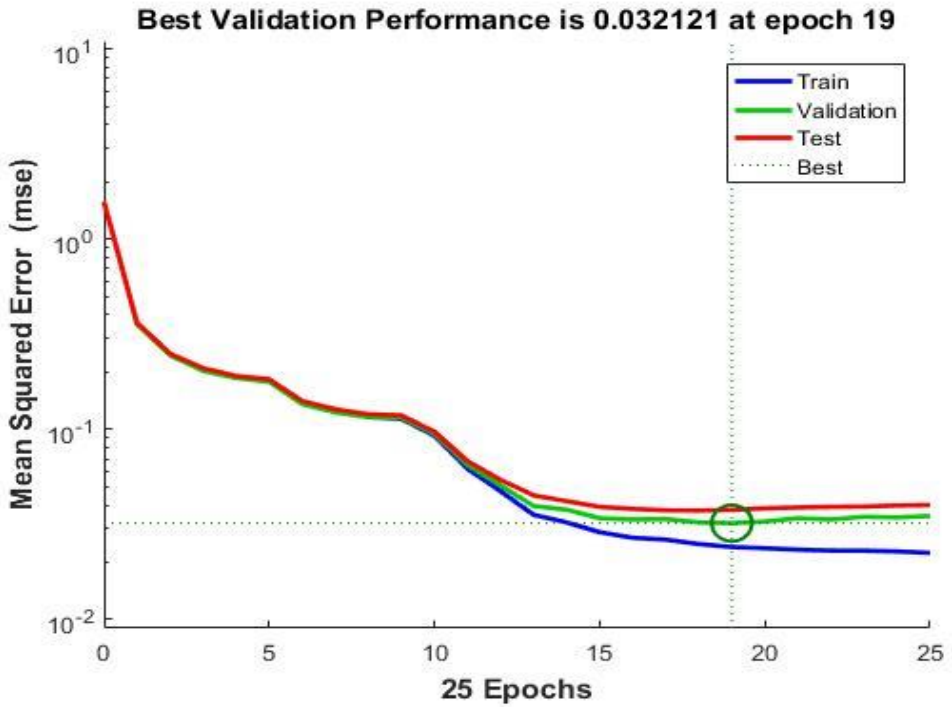


Figure 1. 25 Performance Plot of Experiment Two [Hidden Neurons=20]

Figure 1.26 is the plot of the performance of experiment three for the training, validation, and testing versus the epochs. The blue, green, and red curve is for the training, validation, and testing performance respectively. The dotted black line represents the final goal and the dotted green line represents the best validation performance as a moving line to indicate best validation performance. Best validation performance met was **0.027887 at epoch 15.**

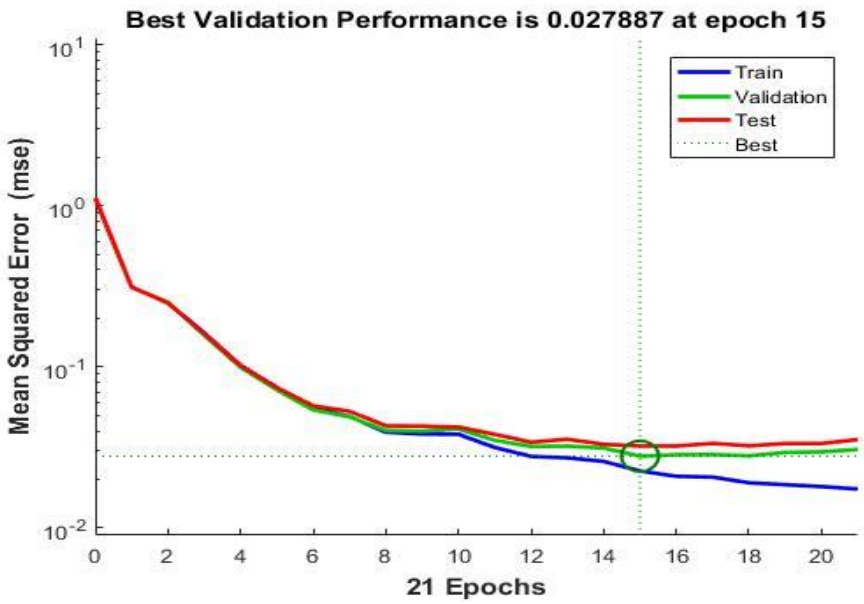


Figure 1. 26 Performance Plot of Experiment Three [Hidden Neurons=25]

4.1.4. Performance Evaluation

Here, the evaluation has been performed in two ways:

- Trained neural network model evaluation based on confusion matrix.
- Real time test using speeches that are prepared for testing.

A. Evaluation Using Confusion Matrix

On the confusion matrix plot, the rows correspond to the predicted class (Output Class) and the columns correspond to the true class (Target Class). The diagonal cells correspond to observations that are correctly identified. The off-diagonal cells correspond to incorrectly identified observations. Both the number of observations and the percentage of the total number of observations are shown in each cell. The column on the far right of the plot shows the percentages of all the examples predicted to belong to each class that are correctly and incorrectly classified. These metrics are often called the **precision** (or positive predictive value) and false discovery rate, respectively. The row at the bottom of the plot shows the percentages of all the examples belonging to each class that are correctly

and incorrectly classified. These metrics are often called the **recall** (or true positive rate) and false negative rate, respectively. The cell in the bottom right of the plot shows the overall accuracy [50].

Neural networks have a confusion matrix for classification problems model. The confusion matrix has 4 possible results in determining the accuracy of the system.

- ❖ True positive [TP] is the correct identification of a speaker.
- ❖ False positive [FP] is the incorrect identification of a speaker,
- ❖ True negative [TN] is the correct identification of the incorrect speaker,
- ❖ False negative [FN] is the incorrect identification of the incorrect speaker.

Currently, accuracy is the most common evaluation parameter used; accuracy measures the classifier's performance by determining the ratio of correct ("true") results over all cases, which are achieved by the classifier.

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

Confusion matrix displays the number of a correct and incorrect prediction made by the model compared with the target classification. The below few points are a description of the confusion matrix variables for multiclass.

- ➢ The total number of test example of any class would be the sum of the corresponding column (i.e. the TP+FN for that class).
- ➢ The total number of FP's for the class is the sum of values in the corresponding row (excluding the TP).
- ➢ The total number of FN's for the class is the sum of values in the corresponding column (excluding the TP).
- ➢ The total number of TN's for a certain class will be the sum of all the columns and rows excluding that class's column and row.

TN=(Summation total samples of class) - (class's sample + class's FP)

- **Recall**

  Recall is a parameter used to evaluate the performance of a classifier and is represented as a ratio of true positives overall actual positives of considered class. This measure provides a better representation of the actual performance of the classifier because the ratio represents only the cases that were actually positive (TP and FN are truly positive).

Recall= TP/ (TP+FN)

- **Precision**

  Precision is predicted positive values calculated as **Precision= TP/ (TP+FP)**

- **F1-score**

  The F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is the harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

F1-score = 2*(precision * recall) / (precision + recall)

Using the above metrics, confusion matrix of the three experiments has been computed and presented below.

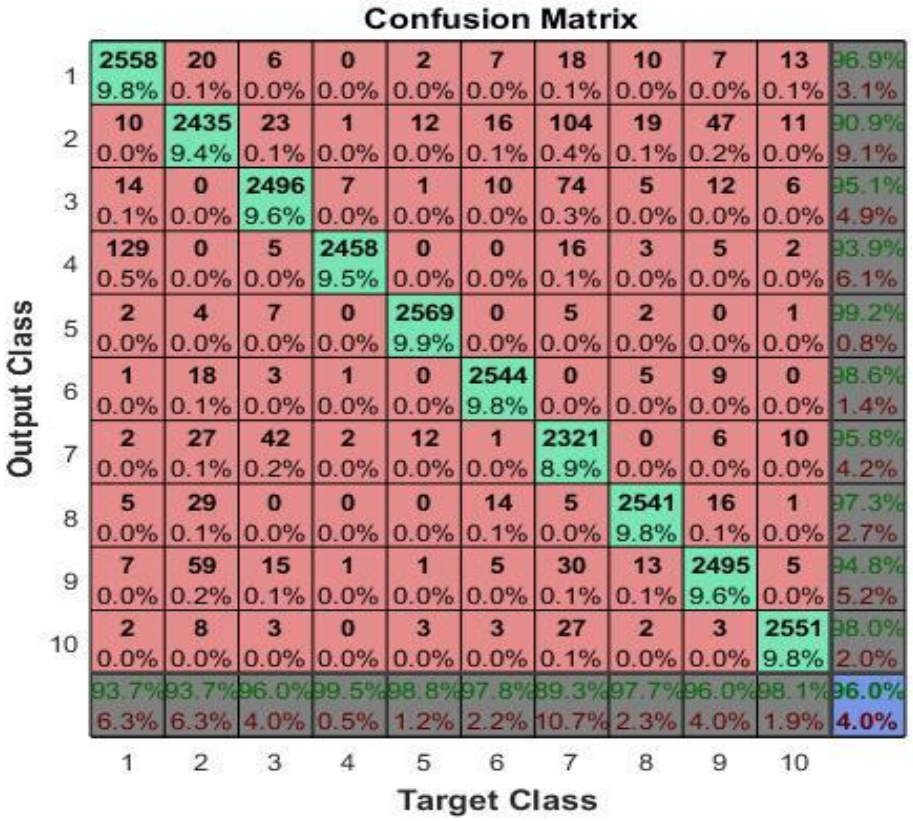


Figure 1. 27 Experiment One: Confusion Matrix [Hidden Neurons=15]

In figure 1.27, the diagonal cells show the number of voice samples that were correctly classified. The blue cell in the bottom right shows the total percent of correctly classified voice samples which is **96.0 %** in the above case and the total percent of wrongly classified voice samples in red which is **4.0%** in the above case. The column on the last right shows the percentages of all the samples predicted to belong to each class that is correctly and incorrectly classified. These metrics are often called the precision (or positive predictive value in green color) and false discovery rate (in red color) respectively. The row at the bottom shows the percentages of how many samples correctly and incorrectly classified respectively from the total samples of the class. For instance, class 1 has 93.7 % [(2558*100)/2730)]. These metrics are often called the recall (or true positive rate) and false negative rate, respectively. The cell in the bottom right of the plot shows the overall accuracy **(96.0%).**

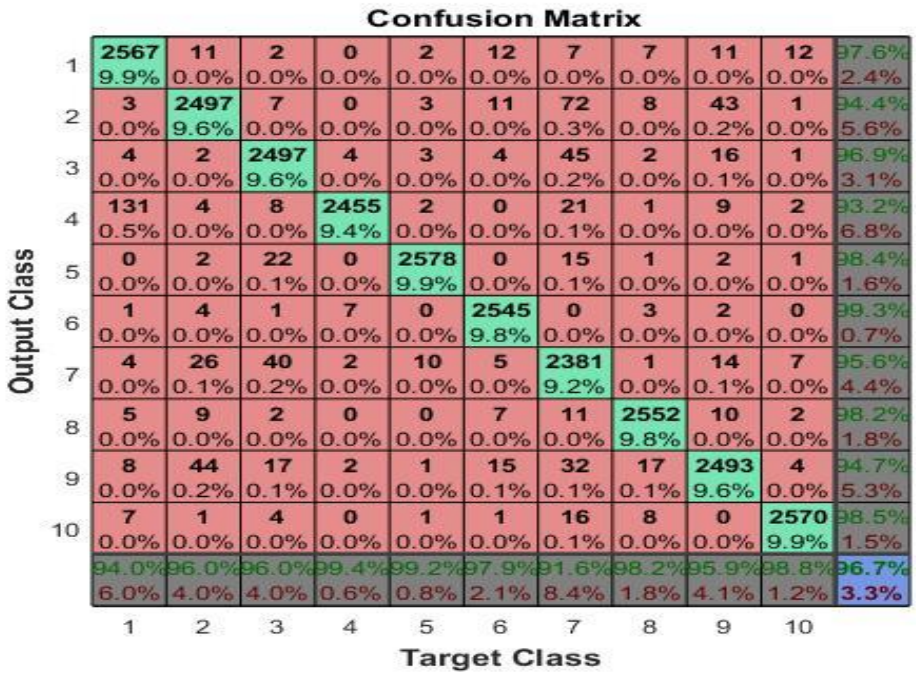


Figure 1. 28 Experiment Two: Confusion Matrix [Hidden Neurons=20]

.

In figure 1.28, the diagonal cells show the number of voice samples that were correctly classified. The blue cell in the bottom right shows the total percent of correctly classified voice samples which is **96.7 %** in the above case and the total percent of wrongly classified voice samples in red which is **3.3%** in the above case. The column on the last right shows the percentages of all the samples predicted to belong to each class that is correctly and incorrectly classified. These metrics are often called the precision (or positive predictive value in green color) and false discovery rate (in red color) respectively. The row at the bottom shows the percentages of how many samples correctly and incorrectly classified respectively from the total samples of the class. For instance, class 1 has 94.0 % [(2567*100)/2730)]. These metrics are often called the recall (or true positive rate) and false negative rate, respectively. The cell in the bottom right of the plot shows the overall accuracy **(96.7%).**

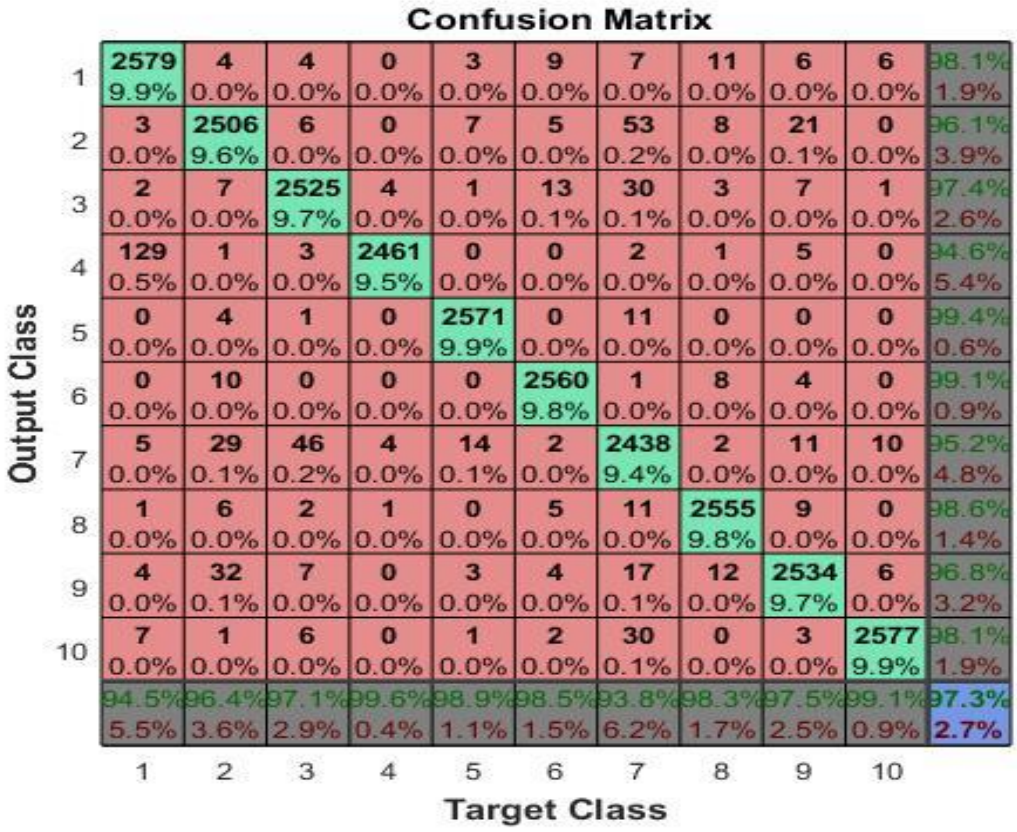


Figure 1. 29 Experiment Three: Confusion Matrix [Hidden Neurons=25]

In figure 1.29, the diagonal cells show the number of voice samples that were correctly classified. The blue cell in the bottom right shows the total percent of correctly classified voice samples which is **97.3 %** in the above case and the total percent of wrongly classified voice samples in red which is **2.7%** in the above case. The column on the last right shows the percentages of all the samples predicted to belong to each class that is correctly and incorrectly classified. These metrics are often called the precision (or positive predictive value in green color) and false discovery rate (in red color) respectively. The row at the bottom shows the percentages of how many samples correctly and incorrectly classified respectively from the total samples of the class. For instance, class 1 has 94.5 % [(2579*100)/2730)]. These metrics are often called the recall (or true positive rate) and false negative rate, respectively. The cell in the bottom right of the plot shows the overall accuracy **(97.3%).**

Based on all confusion matrixes, the below result summary has been summarized.

No. Hidden Neurons	**15**	**20**	**25**
Overall Accuracy	**96.0 %**	**96.7%**	**97.3%**
Male	**49.5%**	**49.8%**	**49.9%**
Female	**50.5%**	**50.2%**	**50.1%**

Table 1. 8 Result Summary Based on Gender

Classes	Hidden Neuron =15					Hidden Neuron =20					Hidden Neuron =25				
	# Samples of Class	**TP**	**FN**	**FP**	**TN**	**# Samples of Class**	**TP**	**FN**	**FP**	**TN**	**# Samples of Class**	**TP**	**FN**	**FP**	**TN**
Class 1	2730	2558	172	83	23214	2730	2567	163	64	23206	2730	2579	151	50	23229
Class 2	2600	2435	165	243	23184	2600	2497	103	148	23252	2599	2506	93	103	23307
Class 3	2600	2496	104	129	23298	2600	2497	103	81	23319	2600	2525	75	68	23341
Class 4	2470	2458	12	160	23397	2470	2455	15	178	23352	2470	2461	9	141	23398
Class 5	2627	2596	31	21	23379	2600	2578	22	43	23357	2600	2571	29	16	23393
Class 6	2600	2544	56	37	23390	2600	2545	55	18	23382	2600	2560	40	23	23386
Class 7	2600	2321	279	102	23325	2600	2381	219	109	23291	2610	2438	172	123	23276
Class 8	2600	2541	59	70	23357	2600	2552	48	46	23354	2600	2555	45	35	23374
Class 9	2600	2495	105	136	23291	2600	2493	107	140	23260	2600	2534	66	85	23324
Class 10	2600	2551	49	51	23376	2600	2570	30	28	23372	2600	2577	23	50	23359
Total	**26027**					**26000**					**26009**				

Table 1. 9 Confusion Matrix Metrics Result of each Class of all Experiments

- **Experiment One:** ***[Hidden Neuron=15]***

Table 1.10 shows how to classify the samples sound frames versus with four confusion matrix parameters. For example, Class 1, has total 2730 sound sample frames stored in sound database extracted from his voice. From those, the truly classified samples are the actual Class 1 predicted correctly 2558 (TP). The 83 (FP) frame samples are classified incorrectly. The 172 (FN) frame samples of actual Class 1 are predicted incorrectly by other classes. The 23214 (TN) frame sampled neither actual class nor predicted class which means truly predicted incorrectly from the total samples **26027**.

- **Experiment Two*: [Hidden Neuron=20]***

Table 1.10 shows how to classify the samples sound frames versus with four confusion matrix parameters. For example, Class 1, has total 2730 sound sample frames stored in sound database extracted from his voice. From those, the truly classified samples are the actual Class 1 predicted correctly 2567 (TP). The 64 (FP) frame samples are classified incorrectly. The 163 (FN) frame samples of actual Class 1 are predicted incorrectly by other classes. The 23206 (TN) frame sampled neither actual class nor predicted class which means truly predicted incorrectly from the total samples **26000**.

- **Experiment Three:** ***[Hidden Neuron=25]***

Table 1.10 shows how to classify the samples sound frames versus with four confusion matrix parameters. For example, Class 1, has total 2730 sound sample frames stored in sound database extracted from his voice. From those, the truly classified samples are the actual Class 1 predicted correctly 2579 (TP). The 50 (FP) frame samples are classified incorrectly. The 151 (FN) frame samples of actual Class 1 are predicted incorrectly by other classes. The 23229 (TN) frame sampled neither actual class nor predicted class which means truly predicted incorrectly from the total samples **26009**.

No. of Hidden Neurons	Metrics	Classes									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
15	**Model Accuracy (%)**	9.8	9.4	9.6	9.5	9.9	9.8	8.9	9.8	9.6	9.8
	Classification Accuracy (%)	99.0	98.4	99.1	99.3	99.8	99.6	98.5	99.5	99.1	99.6
	Precision (%)	96.9	90.9	95.1	93.9	99.2	98.6	95.8	97.3	94.8	98.0
	Recall (%)	93.7	93.7	96.0	99.5	98.8	97.8	89.3	97.7	96.0	98.1
	F1-score (%)	95.3	92.3	95.6	96.6	99.0	98.2	92.4	97.5	95.4	98.1
20	**Model Accuracy (%)**	9.9	9.6	9.6	9.4	9.9	9.8	9.2	9.8	9.6	9.9
	Classification Accuracy (%)	99.1	99.0	99.3	99.3	99.8	99.7	98.7	99.6	99.0	99.8
	Precision (%)	97.6	94.4	96.9	93.2	98.4	99.3	95.6	98.2	94.7	98.5
	Recall (%)	94.0	96.0	96.0	99.4	99.2	97.9	91.6	98.2	95.9	98.8
	F1-score (%)	95.8	95.2	96.5	96.2	98.8	98.6	93.6	98.2	95.3	98.65
25	**Model Accuracy (%)**	9.9	9.6	9.7	9.5	9.9	9.8	9.4	9.8	9.7	9.9
	Classification Accuracy (%)	99.2	99.2	94.5	99.4	99.8	99.8	98.9	99.7	99.4	99.7
	Precision (%)	98.1	96.1	97.4	94.6	99.4	99.1	95.2	98.6	96.8	98.1
	Recall (%)	94.5	96.4	97.1	99.6	98.9	98.5	93.8	98.3	97.5	99.1
	F1-score (%)	96.3	96.3	97.3	97.1	99.2	98.8	94.5	98.5	97.2	98.6

Table 1. 10 Computed Model and Classification Accuracy of all Experiments

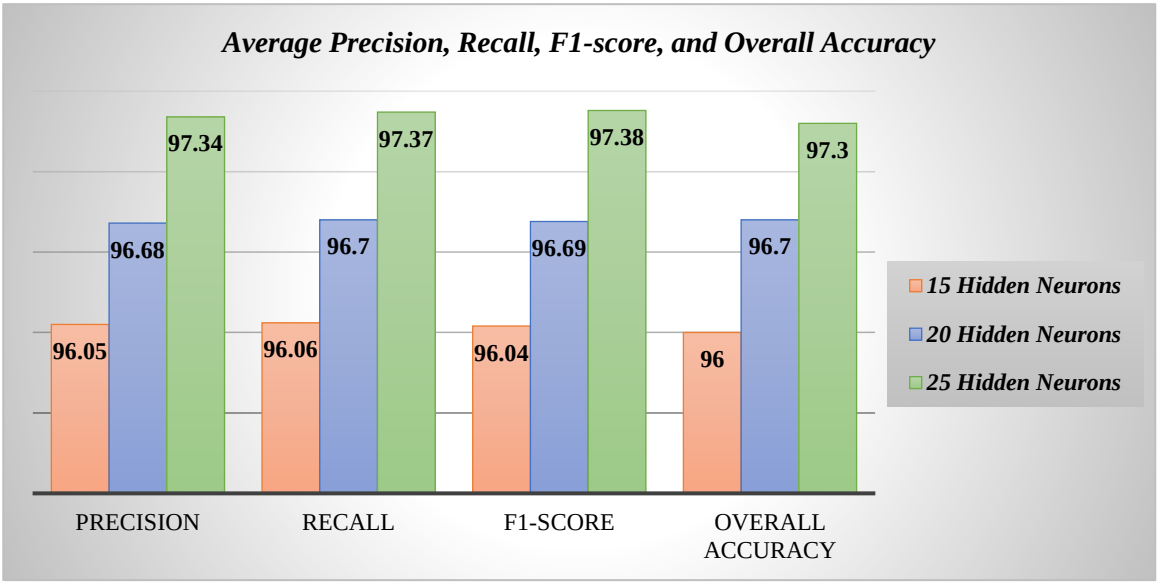


Figure 1. 30 Average Precision, Recall, F1-score, and overall Accuracy of all Experiments

B. Real Time Testing

Here, the prototype has been tested using testing speeches that are preprocessed separately. 20 testing speeches are prepared for each speaker. So, one speaker has been tested 20 times with his/her respective prepared test speeches. ***Appendix D*** shows how the real time testing is conducted. The recognition result is presented below in table 1.12.

Speakers / Experiments	Result	AA	BB	BM	FT	MB	MD	MK	SZ	UAA	YN
Experiment 1	**Correct**	**17**	**17**	**18**	**17**	**18**	**18**	**17**	**18**	**18**	**18**
	Incorrect	**3**	**3**	**2**	**3**	**2**	**2**	**3**	**2**	**2**	**2**
Experiment 2	**Correct**	**18**	**17**	**18**	**19**	**19**	**18**	**17**	**18**	**19**	**19**
	Incorrect	**2**	**3**	**2**	**1**	**1**	**2**	**3**	**2**	**1**	**1**
Experiment 3	**Correct**	**18**	**18**	**19**	**20**	**20**	**19**	**18**	**20**	**18**	**19**
	Incorrect	**2**	**2**	**1**	**0**	**0**	**1**	**2**	**0**	**2**	**1**

Table 1. 11 Real Time Testing Recognition Results

4.2. Discussions

In the experiments and findings section, various findings are presented. The findings have addressed the problems in research question of introduction chapter. The first research question was realizing whether ANN has promising performance for speaker recognition. Definitely, ANN has promising performance for speaker recognition. The indication is separately discussed later on. First, based on the confusion matrix on figure 1.27, 1.28, and 1.29, the overall accuracy is **96.0 %, 96.7%,** and **97.3%** respectively for the three experiments. This shows how often the classifier is correct. Second, based on table 1.10 which presented the performance in terms of TP, TN FP, and FN, the approach showed good results in classifying objects correctly and incorrectly. They were a few objects that are classified wrongly. Third, based on table 1.11 which presented the performance in terms of precision, recall, and F1-score, the approach showed good results. Finally, based on the real time testing on table 1.12, the trained neural network provides a promising result in recognition of untrained speeches of registered speakers.

The implication of the second research question was exploring internal factors that could improve the performance of the model and recognition. Selecting appropriate number of hidden layers and neurons is still a gap. There is no common standard approach, various approaches are listed in[49], from them rule of thumb is used in this study. So, used three different number of hidden neurons [15, 20, and 25] randomly. The findings obtained using this three experiments are presented separately. Figure 1.30 indicated that when the number of hidden number increase the performance also increase in this research context. And based on the three performance plot (figure 1.24, 1.25, and 1.26), the mean squared error is decreased as the number of hidden neuron is increased. Basically, the response times also increases during training the networks when increasing the number of the hidden neurons.

It is not appropriate to compare and contrast this study with the three previously done papers because to compare and contrast, researches should be in common background or equivalent environments. [5] and [6] used 50 and 100 speakers directly recorded speeches respectively, in this study the speeches are collected from multimedia and conducted using 10 speakers. But, the good thing is in [5] , the researcher presented the performance per

number of speakers (10, 20, 30, 40, 50). For 10 speaker, he achieved 83.2% and 87% accuracy using VQ and GMM respectively. In the same manner in [6], the researcher presented the performance per number of speakers (30, 60, 90). For 30 speakers and he achieved 70% and 66% accuracy using VQ and GMM respectively. Therefore, based on this analysis, our study achieved better performance than the two previously worked papers. However, it couldn't mean that this modeling technique has better performance than others for large number of speakers because this study is conducted for only 10 speakers.

At the end, the study introduce a novel attempt of modeling techniques for Amharic language speaker recognition research. And showed that ANN performs better than others on this context.

Author (s) and Year	Title	Feature Extraction Technique	Modeling Technique, Tools	Findings
Aykefam Azene, 2015	*Text-independent speaker identification system for Amharic language.*	MFCC	VQ & GMM, MATLAB	❑ 74.2% & 84.3% accuracy had been achieved for 50 speakers using VQ & GMMs respectively. ❑ For 10 speaker, he achieved 83.2% and 87% accuracy using VQ and GMM respectively.
Abrham Debasu, 2017	*Automatic Text Independent Amharic Language Speaker Recognition in Noisy Environment Using Hybrid Approaches of LPCC, MFCC and GFCC.*	LPCC, MFCC and GFCC.	VQ & GMM, MATLAB	❑ [77.2%, 70.9%, 69%] and [75.2%, 76.9%, 78%] accuracy had been achieved for 30, 60, 90 speakers using VQ and GMMs respectively.
Abrham Debasu, 2017	*Text Independent Amharic Language Dialect Recognition: A Hybrid Approach of VQ and GMM.*	MFCC	VQ & GMM, MATLAB	❑ 85.9% and 92.7% accuracy had been achieved for 25 & 100 speakers respectively.
This Research	*ANN Based Amharic Language Speaker Recognition*	MFCC	ANN, MATLAB	❑ 96%, 96.7%, 97.3% using 15, 20, and 25 number of hidden neurons respectively for 10 speakers.

Table 1. 12 Comparison of this research with other related researches

CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusions

Biometric techniques are one of the modern advances in security systems. No longer requires of entering a password or a PIN which is difficult to remember. Physical characters of the person are used instead. This thesis has presented voiceprint as one of the most promising and useful technologies fitting to the biometric security.

The general issues and applications of speaker recognition is described in the introduction of this thesis in well manner. The goal of the thesis was enabling the environment to develop text-independent speaker recognition for Amharic language using a novel modeling technique which is not attempted for Amharic previously.

Total of 200 sample speeches dataset has been prepared from 10 famous people public speeches. MFCC feature extraction and ANN modeling technique is used to meet the goal. MFCC transformed the preprocessed speech signals in to feature vectors so that it will be appropriate for training the neural network. The feedforward neural network trained the feature vectors using **Levenberg Marquadrt** training algorithm with training epoch parameters of 100. Training speech dataset has been divided into 70%, 15%, 15% for training, validation and testing respectively using **dividerand** function, and **tansig** activation function has been used to convert input signal of a neuron in to output signal.

The study is conducted based on three experiments in changing the number of hidden neurons in to 15, 20, and 25. The experiments have been conducted and the maximum promising findings have been obtained. The proposed modeling technique showed better performance than other techniques which are used in previously done researches.

Findings addressed the whole research questions raised in research question section of introduction chapter. Findings have been discussed in discussion section of chapter four. Since the third research question is general question, the answer is found after completion of the whole research process. Its' implication was exploring external cause that could reduce the performance of the model and recognition accuracy. For the performance improvements of model as well as recognition accuracy, there are various external factors.

From speech perspective;

- ✓ Way of speech collection (direct recorded or from speech database)
- ✓ Environment on which the speech is recorded (Noisy or Noise free)
- ✓ Emotion and health condition of the speaker.
- ✓ Accent of speaker.

From techniques perspective;

- ✓ Appropriateness of preprocessing steps
- ✓ Selecting of best feature extraction technique
- ✓ Modeling using a technique that is best in discriminating.
- ✓ Criteria to evaluate the performance.

Every research context has its' own fitting requirements. It needs scientific searching in order to get an appropriate requirements for specific research context. Still it is a big gap for the researcher to standardize common appropriate requirements that works for any kind of research context. Mimicking the nature is so difficult.

At the end, we witnessed that the architecture selected for the neural network in this thesis, which is a pattern recognition feed-forward neural network with one hidden layer containing 15, 20, and 25 neurons and an output layer containing 10 neurons, inputted 26 coefficients vector was effective and suitable for the identification process.

The study contributes a novel attempt of modeling technique for Amharic language speaker recognition. Since the technique achieved better performance than other modeling techniques that are used in previously worked papers, it gives a clue which modeling technique is best, when anyone who needs to implement practically. The availability of this research will create a chance for the coming researchers who have willingness in this research area in order to conduct comparative analysis on modeling techniques especially for Amharic language speaker recognition. Finally, even though it is not the objective of this research, after all it is the product of this research; the GUI prototype developed by the researcher could be used as a tool for the coming researchers who have willingness in this field of research based on MFCC and ANN. It simplifies searching for different source codes and incorporating all in one.

5.2. Recommendations

As our country is a nation of more than 100 million people with over 80 different languages, lots of attempts concerning speech technology is needed. Research in speech recognition for local language started in 2001. Until now around 40 **speech** recognition researches have been attempted for local languages on different perspectives. Three attempts has been done regarding **speaker** recognition, but they are not enough. In the context of our country, numerous speech and speaker recognition research attempts are needed so as to distinguish the efficiency of approaches and techniques before transforming in to practical. We should attempt many researches with different feature extraction and modeling techniques as much as possible. However, it is a must to find a way to transform the researches in to practical side to side to conducting essential variety researches.

5.3. Future Works

The coming researcher can extend this research by adding the following features.

- Increase number of speakers and sample speeches.
- Including Speaker Verification
- Speaker recognition in noisy environments.
- Hybrid of two or more feature extraction and modeling techniques.
- Amharic dialect recognition.

References

[1] Z. Seifu, "*Hidden Markov Model Based Large Vocabulary, Speaker Independent, Continuous Amharic Speech Recognition*, Unpublished", MSc. Thesis sumitted to School of Graduate Studies Faculty of Informatics, Addis Ababa University", 2003.

[2] S. Furui, "*Pattern Recognition Letters Recent advances in speaker recognition*", Pattern Recognit. Let*t.*, vol. 18, pp. 859–872, 1997.

[3] N. Singh and A. Agrawal, "*Principle and Applications of Speaker Recognition Security System*", no. June 2018.

[4] Solomon Birhanu, "*Isolated Amharic Consonant-Vowel syllable recognition*, Unpublished", MSc. Thesis, Submitted to Addis Ababa University, School of Information Studies for Africa,", 2001.

[5] A. Azene, "*Text-independent speaker identification for the amharic language* ", MSc Thesis submitted to Bahirdar University, no. January, p. 117, 2015.

[6] A. D. Mengistu, "*Automatic Text Independent Amharic Language Speaker Recognition in Noisy Environment Using Hybrid Approaches of LPCC, MFCC, and GFCC*", International Journal of Advanced Studies in Computer Science and Engineering IJASCSE Volume 6 Issue 5, 2017

[7] A. D. Mengistu and D. Melesew, *"Text Independent Amharic Language Dialect Recognition: A Hybrid Approach of VQ and GM*", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 10, no. 1, pp. 215–222, 2017.

[8] S. Singh, "*Forensic and automatic speaker recognition system*", International Journal of Electrical and Computer Enineering., vol. 8, no. 5, pp. 2804–2811, 2018.

[9] Corneliu Octavian Dumitru and Inge Gavat, "*A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language*", 48th International Symposium ELMAR, 2006.

[10] A.P.Henry Charles &G.Devaraj, "Alaigal - A Tamil Speech Recognition", Tamil Internet 2004, Singapore., 2004.

[11] Mohammed Algabri. et al. "*Automatic Speaker Recognition for Mobile Forensic Applications*," Mobile Information Systems, 2017.

[12] S. Ma, Q. Shi, and L. Xu, "*Max–Min Task Scheduling Algorithm for Load Balance in Cloud Computing*", Proceedings of International Conference on Computer Science and Information Technology, Advances in Intelligent Systems and Computing, vol. 255, pp. 341–348, 2014.

[13] Pardeep Sangwan, "*Feature Matching Techniques for Speaker Recognition*", Global Journal of Enterprise Information System, 2018.

[14] Nisha V. S. and M. Jayasheela, "*Survey on Feature Extraction and Matching Techniques for Speaker Recognition Systems*", International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 2, Issue 3, 2013.

[15] beginners guide to speech analysis, accessed 20 May 2019 <https://towardsdatascience.com/beginners-guide-to-speech-analysis.>

[16] Hussien Seid and Bjorn Gamback, "Speaker Independent Continuous Speech Recognizer for Amharic", INTERSPEECH, 2005.

[17] B. Z. Khojah and W. S. Alhalabi, "*Text-Independent Speaker Identification System Using*", Proceedings of The IIER International Conference, no. 1, pp. 22–29, 2016.

[18] General Principles and Applications ,accessed 12 June 2019 <http://www.scholarpedia.org/article/Speaker_recognition#General_Principles_and_Applications >

[19] X. Xue, "*Joint Speech and Speaker Recognition Using Neural Networks*," BSc. Thesis submitted to Electrical Engineering, Novia University of Applied Science, Finland p. 60, 2013.

[20] S. Hamid, "*Speaker Identification System using Wavelets and Neural Networks*,

Unpublished" BSc Thesis. submitted to Department of Electrical and Electronic Engineering Faculty of Engineering and Architecture University of Khartoum no. 044086, 2009.

[21] C. Vimala and V. Radha, "*Suitable Feature Extraction and Speech Recognition Technique for Isolated Tamil Spoken Words*", Internationa. Journal of Computer Science Information Technololgy, vol. 5, no. 1, pp. 378–383, 2014.

[22] F. Kelly, "*Automatic Recognition of Ageing Speakers"*, A dissertation submitted to the University of Dublin," no. April 2014, 2015.

[23] Linear predictive coding, accessed 17 July 2019 <https://en.wikipedia.org/wiki/Linear_predictive_coding.>

[24] Jeet Kumar et al., "*Comparative Analysis of Different Feature Extraction and Classifier Techniques for Speaker Identification Systems : A Review*", International Journal of Innovative Research in Computer and Communication Engineering, pp. 2760–2769, 2014.

[25] N. Dave, "*Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition*", International Journal For Advance Research In Engineering and Technology vol. 1, no. Vi, pp. 1–5, 2013.

[26] K. Kaur and N. Jain, "*Feature Extraction and Classification for Automatic Speaker Recognition System-A Review*," International Journal of Advanced Research Computer Science Software Engineering, vol. 5, no. 1, pp. 1–6, 2015.

[27] J. B. Ramgire and Prof. S. M. Jagdale, "*A Survey on Speaker Recognition With Various Feature Extraction And Classification Techniques*", International Research Journal of Engineering and Technology, Volume: 03 Issue: 04, 2016.

[28] S. Karpagavalli and E. Chandra, "*A Review on Automatic Speech Recognition Architecture and Approaches*", International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol.9, No.4, 2016.

[29] Dr E.Chandra et al., "A *Study on Speaker Recognition System and Pattern classification Techniques*", International Journal of Innovative Research in

Electrical, Electronics, Instrumentation and Control Engineering Vol. 2, Issue 2, 2014

[30] "S. Rajasekaran and G.A. Vijayalakshmi Pai, "*Neural Networks, Fuzzy Logic, and Genetic Algorithms Synthesis and Applications*", Book, 2003.

[31] N. Singh et al., "*Applications of speaker recognition*", International Conference in modeliing, Optimization and Computing, vol. 38, pp. 3122–3126, 2012.

[32] Artificial Neural Networks, accessed 17 February 2019 <https://en.wikipedia.org/wiki/artificial neural networks."

[33] F. G. Beckenkamp, "*A Component Architecture for Artificial Neural Network Systems*", University of Constance, Computer & Information Science, 2002.

[34] An algorithms to train a neural network , accessed 13 July 2019 <https://www.neuraldesigner.com/blog/algorithms_to_train_a_neural_network.>

[35] A. Krenker et al, "*Introduction to the Artificial Neural Networks.*" Book, Published April, 2011

[36] Artificial Neural Networks Applications, accessed 17 February 2019 "https://www.tutorialspoint.com/artificial_neural_network/artificial_neural_network_ applications."

[37] Amharic, accessed 17 January 2019 < https://en.wikipedia.org/wiki/Amharic.>

[38] Tolemariam F.T, "*A typology of verbal derivation in Ethiopian Afro-Asiatic languages*" a dissertation submitted to Leiden University, Netherland, 2009.

[39] Baye Yimam, " የአማርኛሰዋሰዉ. Addis Ababa ት.መ.ማ.ማ. ድ, Unpublished", 2002.

[40] Solomon T and Bantegize A. " *Brana- Application of Amharic Speech Recognition System for Dictation in Judicial Domain*", 2005

[41] Befkadu B. "*Audio-Visual Speech Recognition Using Lip Movement for Amharic Language*. Unpublished" MSc Thesis submitted to Department of Computer Science, Addis Ababa University, 2017

[42] Michael M and Million M, "*Amharic Speech Recognition for Speech Translation*", Atelier Traitement Automatique des Langues Africaines (TALAF), 2016.

[43] S. A. Mahmood and L. E. George, "*Speaker Identification Using Backpropagation Neural Network*," Journal of Zankoy Sulaimani, *December 2008*

[44] M. S. Sinith, et al. "*A novel method for text-independent speaker identification using MFCC and GMM*", International Conference Audio, Language Image Processing, IEEE, 2010.

[45] M. Islam, et al. "*A novel Approach for Text-Independent Speaker Identification Using Artificial Neural Network*", International Journal of Innovative Research and Computer Communication Engineering, vol. 1, no. 4, pp. 838–844, 2013.

[46] Amr Rashed, "*Fast Algorithm for Noisy Speaker Recognition Using ANN*", International Journal of Computer Engineering and Technology (IJCET),Volume 5, Issue 2, 2014.

[47] A. Antony and R. Gopikakumari, "*Speaker identification based on combination of MFCC and UMRT based features*", 8th international conference in adnaces in computing and communication, 2018.

[48] Best audio editing softfware, accessed 20 August 2019 <https://beebom.com/best-audio-editing-software/.>

[49] F. S. Panchal and M. Panchal, "*Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network*", International Journal of Computer Science and Mobile Computing, vol. 3, no. 11, pp. 455–464, 2014.

[50] Plot Confusion, accessed 06 September 2019 <https://www.mathworks.com/help/deeplearning/ref/plotconfusion.html.>

Appendixes

A. Amharic Phonetics

	ä/e [ɛ/ə]	u [u]	i [i]	a [a]	e [e]	ə [ɨ/ʉ]	o [o/ɔ]
h [h]	ሀ ha	ሁ hu	ሂ hi	ሃ ha	ሄ he	ህ hə	ሆ ho
l [l]	ለ lä	ሉ lu	ሊ li	ላ la	ሌ le	ል lə	ሎ lo
ḥ [h]	ሐ ḥa	ሑ ḥu	ሒ ḥi	ሓ ḥa	ሔ ḥe	ሕ ḥə	ሖ ḥo
m [m]	መ mä	ሙ mu	ሚ mi	ማ ma	ሜ me	ም mə	ሞ mo
ś [s]	ሠ śä	ሡ śu	ሢ śi	ሣ śa	ሤ śe	ሥ śə	ሦ śo
r [r]	ረ rä	ሩ ru	ሪ ri	ራ ra	ሬ re	ር rə	ሮ ro
s [s]	ሰ sä	ሱ su	ሲ si	ሳ sa	ሴ se	ስ sə	ሶ so
š [ʃ]	ሸ šä	ሹ šu	ሺ ši	ሻ ša	ሼ še	ሽ šə	ሾ šo
q [kʾ]	ቀ qä	ቁ qu	ቂ qi	ቃ qa	ቄ qe	ቅ qə	ቆ qo
b [b]	በ bä	ቡ bu	ቢ bi	ባ ba	ቤ be	ብ bə	ቦ bo
v [v]	ቨ vä	ቩ vu	ቪ vi	ቫ va	ቬ ve	ቭ və	ቮ vo
t [t]	ተ tä	ቱ tu	ቲ ti	ታ ta	ቴ te	ት tə	ቶ to
č [ʧ]	ቸ čä	ቹču	ቺ či	ቻ ča	ቼ če	ች čə	ቾ čo
ḫ [h]	ኀ ḫa	ኁ ḫu	ኂ ḫi	ኃ ḫa	ኄ ḫe	ኅ ḫə	ኆ ḫo
n [n]	ነ nä	ኑ nu	ኒ ni	ና na	ኔ ne	ን nə	ኖ no
ny/ñ [ɲ]	ኘ nyä	ኙ nyu	ኚ nyi	ኛ nya	ኜ nye	ኝ nyə	ኞ nyo
ʾ/ʔ [ʔ]	አ ʾa	ኡ ʾu	ኢ ʾi	ኣ ʾa	ኤ ʾe	እ ʾə	ኦ ʾo
	ä/e [ɛ/ə]	u [u]	i [i]	a [a]	e [e]	ə [ɨ/ʉ]	o [o/ɔ]

	ä/e [ɛ/ə]	u [u]	i [i]	a [a]	e [e]	ə [ɨ/ʉ]	o [o]
k [k]	ከ kä	ኩ ku	ኪ ki	ካ ka	ኬ ke	ክ kə	ኮ ko
ķ [ḳ]	ኸ ķä	ኹ ķu	ኺ ķi	ኻ ķa	ኼ ķe	ኽ ķə	ኾ ķo
w [w]	ወ wä	ዉ wu	ዊ wi	ዋ wa	ዌ we	ው wə	ዎ wo
ʿ/ʿ [ʔ]	ዐ ʿa	ዑ ʿu	ዒ ʿi	ዓ ʿa	ዔ ʿe	ዕ ʿə	ዖ ʿo
z [z]	ዘ zä	ዙ zu	ዚ zi	ዛ za	ዜ ze	ዝ zə	ዞ zo
ž [ʒ]	ዠ žä	ዡ žu	ዢ ži	ዣ ža	ዤ že	ዥ žə	ዦ žo
y [j]	የ jä	ዩ ju	ዪ ji	ያ ja	ዬ je	ይ jə	ዮ jo
d [d]	ደ dä	ዱ du	ዲ di	ዳ da	ዴ de	ድ də	ዶ do
ǧ [ʤ]	ጀ ǧä	ጁ ǧu	ጂ ǧi	ጃ ǧa	ጄ ǧe	ጅ ǧə	ጆ ǧo
g [g]	ገ gä	ጉ gu	ጊ gi	ጋ ga	ጌ ge	ግ gə	ጎ go
ṭ [tʾ]	ጠ ṭä	ጡ ṭu	ጢ ṭi	ጣ ṭa	ጤ ṭe	ጥ ṭə	ጦ ṭo
č [ʧʾ]	ጨ čä	ጩ ču	ጪ či	ጫ ča	ጬ če	ጭ čə	ጮ čo
p [pʾ]	ጰ pä	ጱ pu	ጲ pi	ጳ pa	ጴ pe	ጵ pə	ጶ po
ṣ [ts]	ጸ ṣä	ጹ ṣu	ጺ ṣi	ጻ ṣa	ጼ ṣe	ጽ ṣə	ጾ ṣo
ṣ́ [ts]	ፀ ṣ́ä	ፁ ṣ́u	ፂ ṣ́i	ፃ ṣ́a	ፄ ṣ́e	ፅ ṣ́ə	ፆ ṣ́o
f [f]	ፈ fä	ፉ fu	ፊ fi	ፋ fa	ፌ fe	ፍ fə	ፎ fo
p [p]	ፐ pä	ፑ pu	ፒ pi	ፓ pa	ፔ pe	ፕ pə	ፖ po
	ä/e [ɛ/ə]	u [u]	i [i]	a [a]	e [e]	ə [ɨ/ʉ]	o [o]

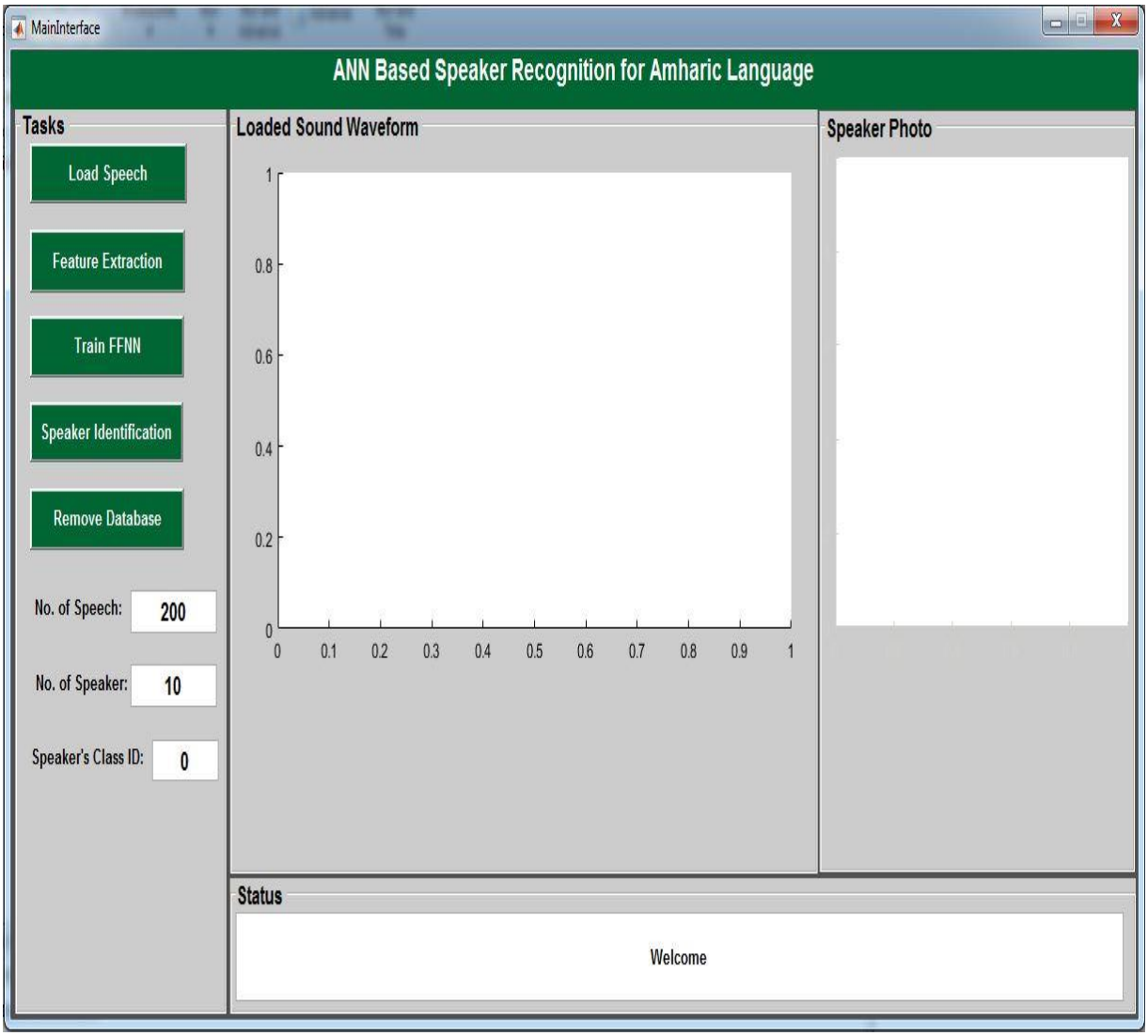
[lʷa]	[hʷa]	[mʷa]	[sʷa]	[rʷa]	[sʷa]	[ʃʷa]	[kʷə]	[kʾʷi]	
[kʾʷa]	[kʾʷc]	[kʾʷi/ʉ]	[bʷa]	[vʷa]	[tʷa]	[ʧʷa]	[hʷə]	[hʷi]	
[hʷa]	[hʷc]	[hʷi/ʉ]	[nʷa]	[ɲʷa]	[ʔa]	[kʷə]	[kʷi]	[kʷa]	
[kʷc]	[kʷi/ʉ]	[hʷə]	[hʷi]	[hʷa]	[hʷc]	[hʷi/ʉ]	[zʷa]	[ʒʷa]	
[tsʾʷa]	[ʤʷa]	[gʷə]	[gʷi]	[gʷa]	[gʷc]	[gʷi/ʉ]	[tʾʷa]	[ʧʾʷa]	
[pʾʷa]	[tsʷa]	[fʷa]	[pʷa]						

B. Sample code [createnn.m]

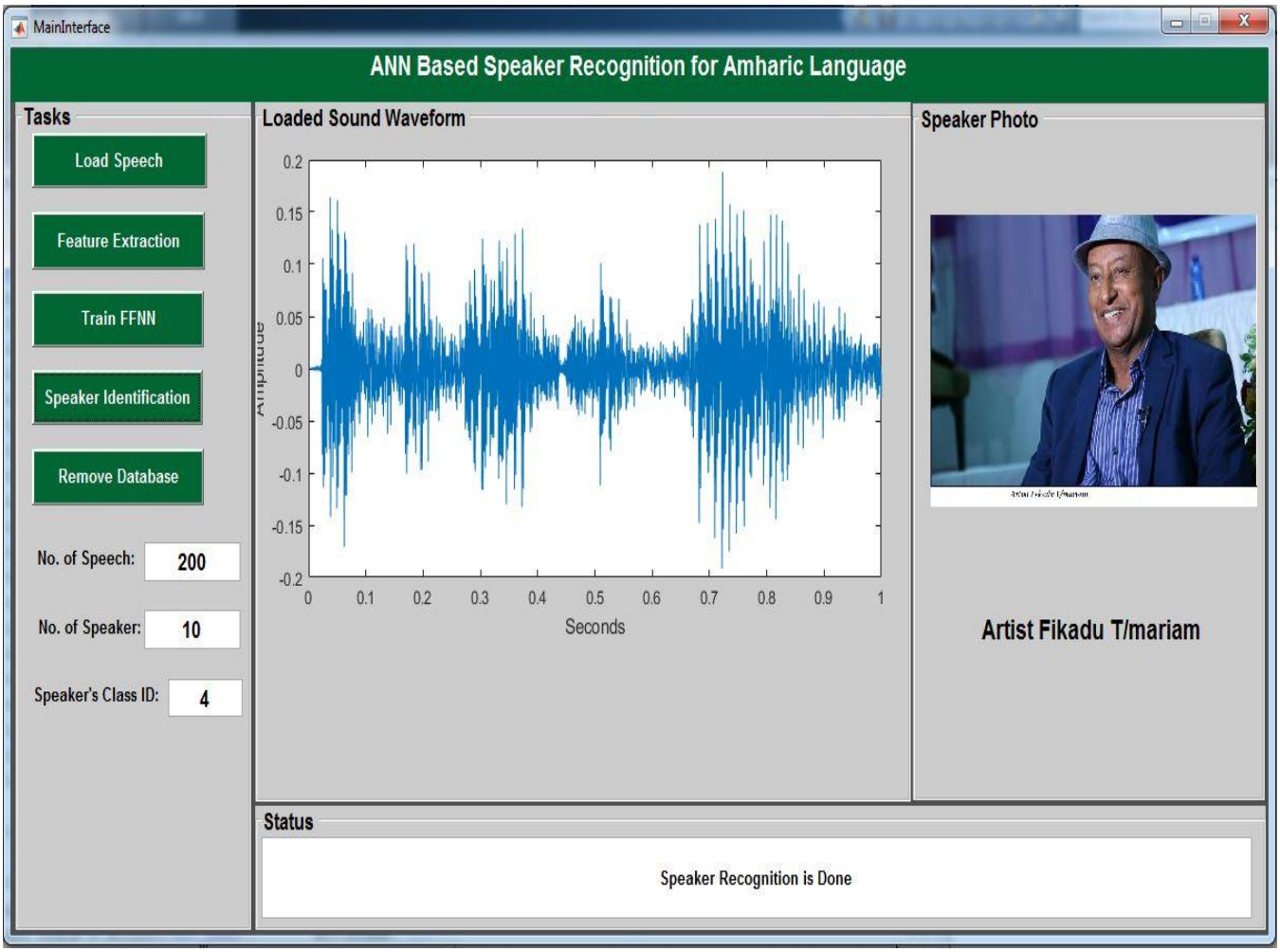
```
   function [MyNetwork] = createnn()
   %inputs = P;
   %targets  = T;
   %*****************************************************************
   ******
   load('speech_feature_vectors_database.dat','-mat');
   P = [];
   T = [];
   L = features_size;
   for i=1:L
       fmatrix = features_data{i,1};
       [dimx,dimy] = size(fmatrix);
       P = [P fmatrix'];
       t   = zeros(max_class-1,dimx);
       pos =  features_data{i,2};
       for j=1:(max_class-1)
           if j==pos
               t(j,:) = 1;
           else
               t(j,:) = 0;
           end
    end
       T = [T t];
   end

   % CREATE NEURAL NETWORK

MyNetwork=feedforwardnet(15,'trainlm');
   MyNetwork.layers{1}.transferFcn = 'tansig';
MyNetwork.layers{2}.transferFcn = 'tansig';

   MyNetwork.LW{1,1}           = MyNetwork.LW{1,1}*0.01;
   MyNetwork.b{1}              = MyNetwork.b{1}*0.01;
   MyNetwork.performFcn        = 'mse';
   MyNetwork.trainParam.goal   = 0;  % error goal
   MyNetwork.trainParam.show   = 10;    %epochs showing intervals
   MyNetwork.trainParam.epochs = 1000;  % maximum iterations
   MyNetwork.trainParam.mc     = 0.95;    % Momentum constant
   MyNetwork.trainParam.lr=0.01; %learning rate
   %*****************************************************************
   ****
   MyNetwork.divideFcn='dividerand';
   MyNetwork.divideMode = 'sample'; % Divide up every sample
MyNetwork.divideParam.trainRatio = 70/100;%Training Ratio
   MyNetwork.divideParam.valRatio = 15/100;%Validation Ratio
   MyNetwork.divideParam.testRatio = 15/100;%Test Ratio
   MyNetwork.input.processFcns = {'removeconstantrows','mapminmax'};
   MyNetwork.output.processFcns = {'removeconstantrows','mapminmax'};
   %*****************************************************************
   ****
   MyNetwork.trainParam.showWindow=1; % Show training GUI
```

```
MyNetwork.trainFcn = 'trainlm';
%P                      = inputs;
%T                      = targets;
MyNetwork = init(MyNetwork);
[MyNetwork,tr]              = train(MyNetwork,P,T);

save MyNetwork;
y=MyNetwork(P);

figure, plotconfusion(T,y)
%-------------------------------------------------
```

C. GUI Interface of the Prototype



D. GUI Result given in Real Time Testing



E. Glossary

Terms	Definition
ANN Architecture	A description of the number of the layers in a neural network, each layer's transfer function, the number of neurons per layer, and the connections between layers.
ANN Topology	The way the neurons are connected, and it is an important factor in network functioning and learning.
Articulation	Is the act of expressing something in a coherent verbal form, or an aspect of pronunciation involving the articulatory organs
Authentication	The process of determining whether someone or something is, in fact, who or what it declares itself to be.
Backpropagation	iterative, recursive and efficient algorithm for calculating the weights updates to improve the network until it is able to perform the task for which it is being trained
Band-pass filter	A device that passes frequencies within a certain range and rejects (attenuates) frequencies outside that range.
Bias	A neuron parameter that is summed with the neuron's weighted inputs and passed through the neuron's transfer function to generate the neuron's output.
Biometrics	The use of the measurement of human physiological or behavioral characteristics for the purpose of identifying individuals or verifying their claims.
Bit depth	Best thought of as a container for the amplitude of audio.
Classification	An association of an input vector with a particular target vector.
Confusion Matrix	a simple way to layout how many predicted categories or classes were correctly predicted and how many were not
Connection	A one-way link between neurons in a network.

Connection Strength	The strength of a link between two neurons in a network. The strength, often called weight, determines the effect that one neuron has on another.
Cycle	A single presentation of an input vector, calculation of output, and new weights and biases.
Epoch	The presentation of the set of training (input and/or target) vectors to a network and the calculation of new weights and biases.
Feature Vector	A vector that contains unique information describing a speakers' speech important characteristics.
Feedforward Network	A layered network in which each layer only receives inputs from previous layers.
Forensic Investigation	The gathering and analysis of all crime-related physical evidence in order to come to a conclusion about a suspect.
Frequency Domain	The analysis of mathematical functions or signals with respect to frequency, rather than time (graph shows how much of the signal lies within each given frequency band over a range of frequencies.)
Gradient Descent	The process of making changes to weights and biases, where the changes are proportional to the derivatives of network error with respect to those weights and biases.
Input Layer	A layer of neurons receiving inputs directly from outside the network.
Input Vector	A vector presented to the network.
Input Weight Vector	The row vector of weights going to a neuron
Input Weights	The weights connecting network inputs to layers
Learning	The process by which weights and biases are adjusted to achieve some desired network behavior.
Learning Rate	A training parameter that controls the size of weight and bias changes during learning.
Linear Activation Function	A transfer function that produces its input as its output.
Mean Squared Error (MSE)	The average squared error between the networks outputs **O** and the target outputs **T**. (difference between the estimated values and the actual value.)

Mean Squared Error Function	The performance function that calculates MSE
Momentum	A technique often used to make it less likely for a backpropagation networks to get caught in a shallow minima.
Momentum Constant	A training parameter that controls how much "momentum" is used.
Neuron	The basic processing element of a neural network. Includes weights and bias, a summing junction and an output transfer function.
Non-stationary	Statistical properties change over time.
Output Layer	A layer whose output is passed to the world outside the network
Output Vector	The output of a neural network. Each element of the output vector is the output of a neuron.
Output Weight Vector	The column vector of weights coming from a neuron or input.
Pattern Recognition	The task performed by a network trained to respond when an input vector close to a learned vector is presented. The network "recognizes" the input as one of the original target vectors
Phonetics	a branch of linguistics that studies the sounds of human speech, or—in the case of sign languages—the equivalent aspects of sign
Principal Component Analysis (PCA)	Orthogonalize the components of network input vectors. This procedure can also reduce the dimension of the input vectors by eliminating redundant components.
Prototype	an early sample, model, or release of a product built to test a concept or process
Sampling rate	Best thought of as a container for frequency information.
Soft-Computing	Tolerant of impression, uncertainty, partial truth, and approximation. In effect, the role model for soft computation is the human mind.
Speech Corpus	is a database of speech audio files and text transcriptions
Stationary	Statistical properties such as the mean, variance and autocorrelation are all constant over time.

Surveillance	Continuous gathering of health data needed to monitor the population's health status in order to provide or revise needed services.
Target Vector	The desired output vector for a given input vector.
Test Vectors	A set of input vectors (not used directly in training) that is used to test the trained network.
Time Domain	The analysis of mathematical functions, physical signals or time series of economic or environmental data, with respect to time. (graph shows how a signal changes over time)
Training	The process of modifying the weights in the connections between network layers with the objective of achieving the expected output is called training a network.
Training Vector	An input and/or target vector used to train a network.
Under resourced Language	a language which has poor usage in technology
Update	Make a change in weights and biases.
Validation Vector	A set of input vectors (not used directly in training) that is used to monitor training progress so as to keep the network from overfitting.
Weight functions	Apply weights to an input to get weighted inputs as specified by a particular function.
weighted input vector	The result of applying a weight to a layer's input, whether it is a network input or the output of another layer.
Rule of Thumb	a principle with broad application that is not intended to be strictly accurate or reliable for every situation

F. የመመረቂያ ፅሁፉ ባለቤት ምክሮች

ከታች የተዘረዘሩት ምክረ ሃሳቦች የመመረቂያ ፅሁፉ ባለቤት ይህን ጥናታዊ ፅሁፍ ሲሰራ ከጅምሩ እስከ ፍፃሜው ያለፋቸውን ተሞክሮዎች ሲሆኑ ለአዳዲስ የመመረቂያ ፅሁፉ ሰሪዎች ለቅድመ ዝግጅት እገዛ እንዲሆን በማሰብ ነው::

የመመረቂያ ፅሁፍ (Thesis)- ማለት ስለ አንድ ርዕሰ ጉዳይ ችግሮች ማየት እስኪቻል ድረስ በጥልቀት ማወቅ እና ችግሮችን መለየት አንዲሁም ለተገኙት ችግሮች የመፍትሔ ሃሳብ እስከ መጠቆም ድረስ ያጠቃልላል:: የመመረቂያ ፅሁፍ የራሱ ሳይንሳዊ የሆነ አሰራር፣ አፃፃፍና አቀራረብ አለው:: ሳይንሳዊ ቅደም ተከተሉን ጠብቆ መስራት ጥሩ ውጤት ከሚያስገኙ ምክንያቶች አንዱ ነው::

1. አንድ የመመረቂያ ፅሁፉ ሰሪ በቅድሚያ የመመረቂያ ፅሁፍ የሚሰራበትን መስክ (Thematic Area) መምረጥ ይጠበቅበታል:: የመመረቂያ ፅሁፍ Thematic Area ከዚህ በፊት ከወሰዳቸው ወይም እየወሰደ ካላቸው ኮርሶች ጋር ተያያዥነት ቢኖረው ይመረጣል:: ምክንያቱም ሂደቱን ያቀልለታል:: ቢችል የማስተርስ ፕሮግራም መማር ከመጀመሩ በፊት Thematic Area መርጦ ዝግጅት ቢያደርግ በጣም ይቀለዋል:: የመመረቂያ ፅሁፉ መስራት ከመጀመሩ በፊት በቂ እውቀት እንዳለውና ሊጨርሰው መቻሉን ራሱን ሳይዋሽ ማረጋገጥ ይጠበቅበታል:: መሃል ላይ ርዕስ መቀየር የማይችልበት ምዕራፍ ላይ ከደረሰ ተስፋ ሊያስቆርጥ ይችላል:: ስለዚህ ከፍተኛ የማያዳግም ጥንቃቄ ለርዕስ መረጣ
2. ርዕሱ ተቀባይነት እንዲያገኝ አሳማኝ የሆነ Proposal ሳይንሳዊ በሆነ መንገድ ማዘጋጀት
3. ርዕሱ ተቀባይነት ካገኘ በኋላ ስለርዕሱና ከርዕሱ ጋር ግንኙነት ስላላቸው ሃሳቦች በቂ resource material መሰብሰብ (Books, PPT, Video, Papers) እና በኮምፒተርህ በቀላሉ ማግኘት በምትችልበት አቀማመጥ ማስቀመጥ። የሆነ ሰዓት resource download ማድረግ ማቆም መወሰን አለብህ:: ሁሉም የሚጠቅም ስለሚመስልህ በተደጋጋሚ resource download የማድረግ አባዜ ውስጥ ልትገባ ትችላለህ::

4. ለረጅም ጊዜ ስለርዕሱና ከርዕሱ ጋር ግንኙነት ስላላቸው ሃሳቦች በጥልቀት አንብብ (የመመረቂያ ፅሁፍህን ለማጠናቀቅ ካለህ ጊዜ ውስጥ 70%ን አንብብ) ሙሉ እውቀት ከያዝክ በኋላ ለimplementation አንድ ወር እንዲሁም ዶክመንቱን (Thesis Report) ለማዘጋጀት አንድ ወር በጣም በቂ ነው:: ስለርዕስ ጉዳይህ ስታነብ ከስር መሰረቱ ለማንበብ ሞክር በተለይ መፅሃፍ ብታነብ ይመከራል:: PPT እና Video እንደማጠናከሪያ መጠቀም ጥሩ ነው::
5. ከአማካሪህ (Advisor) ጋር ጥብቅና ጥሩ የሆነ ግንኙነት ለመፍጠር ሞክር ምክንያቱም ከአንተ በልምድ ይበልጣሉና መንገድ ያሳዩሃል:: በመመረቂያ ፅሁፍህ ላይ ለመቀየር ስላሰብካቸው ነገሮች በሙሉ ለአማካሪህ ማሳወቅህን በፍፁም እንዳትረሳ
6. ከአንተ ርዕስ ጋር ተቀራራቢ የሆነ ስራ ከሰሩ ሰዎች ጋር የማውራት ልምድ ቢኖርህም አሪፍ ነው:: ከሃገር ውጪ ከሆኑም በኢሜል እስከማውራት ድረስ ማለት ነው::
7. ዶክመንቱን ስታዘጋጅ ትኩረት ማድረግ ይጠበቅብሃል ምክንያቱም 70% ነጥብ ስለሚይዝ:: ስለዚህ እንዴት ቢፃፍ ጥሩ እንደሚሆን ሊያስረዱ የሚችሉ መፅሐፎች አሉ እነሱን አንብብ:: እያንዳንዱ ምዕራፍ የራሱ የሆነ ይዘት አለው::ዶክመንትህ ውስጥ አንተ ያለገባህ አንድም ሃሳብ መኖር የለበትም::
8. ለDefense ስራዬን ለመግለፅ ይመቻሉ የምትላቸውን ሃሳቦች ብቻ በPPT አካት:: በተቻለ መጠን PPT ላይ ፅሁፍ አታብዛ ሃሳቦችህን በስዕል ለመግለፅ ሞክር:: PPT ለዕይታ ማራኪ እንዲሆን ጥረት አድርግ:: ያዘጋጀኸውን PPT ደጋግመህ መለማመድ ብትችል በጣም ጥሩ ነው:: በራስ መተማመንህን ከፍ ያደርገዋል::