ADAMA SCIENCE & TECHNOLOGY UNIVERSITY

SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTING

DEPARTMENT OF COMPUTING

A MASTER'S THESIS

ON

TEXT DEPENDENT SPEAKER RECOGNITION USING ARTIFICIAL NEURAL NETWORKS FOR TIGRIGNA LANGUAGE UTTERANCE

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF

MASTER OF SCIENCE IN INFORMATION SYSTEMS

BY:

ATAKLTI NGUSE

Advisor: Mohammed Wazih (Assistant Professor)

June 2018

Adama, Ethiopia

Dedication

I dedicate this MSc thesis to my MOM, Brother, Sister and to all people who have had their own share of contribution throughout my life. Love you!

Table of Contents **page numbers**

List of Tables

List of Figures

List of Abbreviation

Abbreviation / Acronym	Meaning
ADC	Analog to Digital Convertor
ANNs	Artificial Neural Networks
BP	Back Propagation
CV/CVC	Consonant Vowel /Consonant Vowel Consonant
DCT	Discrete Cosine Transform
D.N.A	Deoxyribonucleic Acid
DTW	Dynamic Time Warping
EPD	End Point Detection
FFNN	Feed Forward Neural Network
FFT	Fast Fourier Transform
FN	False Negative
FNR	False Negative Rate
FP	False Positive
FPR	False Positive Rate
GMM	Gaussian Mixture Mode
GUI	Graphical User Interface
HCI	Human Computer Interaction
HMM	Hidden Markov Model
Hz	Hertz
ID	Identification
IPA	International Phonetic Alphabet
LDC	Linguistic Data Consortium
LPC	Linear Predictive Coding
LPCC	Linear Prediction Cepstrum Coding
MATLAB	Matrix Laboratory
MFCC	Mel-Frequency Cepstrum Coefficients
MLP	Multi-Layer Perceptron
NIST	National Institute of Standards and Technology

NN	Neural Network
PC	Personal Computer
PIN	Personal Identification Number
ROC	Receive Operating Characteristics
SVM	Support Vector Machine
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
VQ	Vector Quantization

Acknowledgment

First and foremost, I would like to thanks the almighty God, who gave me the strength to sit up nights and the knowledge to complete and precede the thesis successfully.

Second, I extend my sincere gratitude to my thesis advisor, Mohammed Wazih (Assistance professor) abundance of patience. He never lost faith in my thesis when I faltered and got stuck. His continued guidance has taught me a valuable lesson regarding research methods.

I sincerely acknowledge for those they were willing to record all speakers and the entire researcher for their wide contribution in the area of speaker recognition technology that gives us a ray of light to understand and work for further improvement in the field of speaker recognition.

I would like to thank everyone whose help, support and encouragement have helped me through the completion of this thesis.

Finally, I would like to thank my family for their continued prayers and support through this entire period of study. I dedicate this thesis to them and their faith in me.

Abstract

This paper attempted to implement a text-dependent speaker recognition system based on the artificial neural network for Tigrigna language. Speaker recognition is one of the biometric technologies that have been developed almost more than three decades ago. Speaker recognition is a computing task which discriminates between people based upon their voice characteristics. Speaker recognition has divided into two categories based on its tasks: speaker identification and speaker verification. Speaker identification is the process of determining unknown speaker from a known group of speakers. Meanwhile, speaker verification is the process of accepting or rejecting the identity claim of a speaker. Speaker recognition for different languages is still a big challenge for researchers in terms of identification rate accuracy and securities are among the main issues. So, it's vital to find out the root cause of the accuracy performance gap for the speaker model. With the advanced technologies that have been achieved today, researchers have found that using biometric voice as the security system of the efficient one method.

The specific activity of speaker recognition methodologies has been used in this thesis. The method divided into two parts. The first part is training process to build a reference model and the second part is testing for the identification process. For identification process, the sound corpus is collected from different speakers considering the terms gender and age taken at sampling frequency 44.1 kHz and16 bit. The development steps of the speaker recognition prototype were dividing into three modules. At first, speech pre-processing techniques such as endpoint detection applied to remove the silent speech. Then, speaker voice characteristics can be extracted using Mel-frequency cepstrum coefficients form feature vector and stored in the database as a reference model. At last, at the stage of pattern matching the unknown inputted speech and the stored reference model are compared using neural network model for the purpose of training the model and identifying the speaker.

Performance of the model is measured by parameters of the confusion matrix. The accuracy of the model for text-dependent speaker recognition MFCC based on MLP NN for Tigrigna language is 96.6% and 93.7% for the word and phrase respectively tested by 10 speakers. ANN speaker model showed good promising results. All the tasks were implemented using MATLAB.

Keywords: Artificial neural networks, Speaker recognition, Mel-Frequency cepstrum coefficient, Feed-forward neural networks, Multilayer Perceptron

CHAPTER ONE

1. THE PROBLEM AND ITS BACKGROUND

1.1. Introduction

Since the 1960s, computer scientists have been looking research techniques for making computer able to record, interpret and understand human speech. Even the most enormous problem such as sampling voice was a huge challenge in the early years [13].

In the past six decades, machine learning has been providing new solutions for extremely abstract tasks. Such as, adaptation of human brain biological nervous system to modern digital computer systems. Numerous advances have been made in developing intelligent systems for digital computers make tasks easy; Researchers from some inspired by many scientific disciplines biological neural are designing networks for AI to adapt machine learning. Among many of machine learning is artificial neural networks. It solves a variety of problems in pattern recognition, prediction, optimization, associative memory, and motor control [3].

As part of pattern recognition as solution computer-based speech recognition speech in one of the main naturalness and ease of use of speech that forms the primary of motivating factor behind of speaker recognition. Because of speech signal conveys two main levels of information first carries the message and reveals speakers' information like language being spoken, speaker emotions, gender and identity of the speaker [1]. The main objective of automatic speaker recognition is to extract, characterize and recognize the information about speakers' identity. Since every person has a unique feature in his/her voice, it is valuable to classify between two persons based their own voice.

Voices will change in volume, speed, pitch, tone and other different aspects. Because of various social foundations voices may differ in terms of accent, articulation, and pronunciation. Such kinds of elements make a difference in the implementation of speech and speaker recognition a very challenging problem.

Biometrics is an emerging technology based on pattern recognition system for automatically identifying individuals using distinct physical or behavioral characteristics. In today's society, people prefer options in technology that simplify activities. The aim of these technological

efforts is to produce completely automated systems for personal identification or verification that are convenient to use and offer high performance and reliability. Using speech recognition, users could type text or issue device commands through a speech by their voices. This kind of systems like language translation and dictation may become easy and simplified hands-free devices [5].

The communications between humans and computers are wider and deeper by communication functions using a mouse, keyboard, and touchscreen cannot satisfy the fast, accurate and efficient interchange of information. How to send information in a more natural, more efficient and quicker way has become an urgent question based on speech [2].

Speech processing is nowadays as one of the important application and research area of digital signal processing. Various fields for research in speech processing are speech recognition, speaker recognition, speech synthesis, and speech coding. Speaker recognition is the process of automatically extracting personal identity information by analysis of spoken utterances, and is different from speech recognition of which purpose is to recognize the content of the speech. Speaker recognition involves two major tasks: verification and identification [4]. Identification refers to determining the identities of who is speaking while verification is a process of claimant accept or reject based on the binary decision.

This research has been done on text-dependent speaker recognition by using artificial neural network model for Tigrigna language utterance. The language focus in this research is Tigrigna which is one of the under resourced Ethiopian Semitic languages and also spoken in Tigray, Eritrea and some part of Ethiopia by dialects of Tigrigna.

In addition, the researcher intended for growing one step towards contribution on the field of voice recognition for Tigrigna utters by adding knowledge to the state of the art. It could be a motivation and base for other researchers to will conduct in the various area of Ethiopian language. As we known Ethiopia is rich in languages which mean it has a multilingual country.

Hence, as the researcher, there are not known yet attempts in the area of text-dependent speaker recognition for Tigrigna language so far. Therefore, this study aimed at investigating and proving out the possibility of developing a prototype speaker dependent for native Tigrigna speaker recognition system based on artificial neural network (ANN). The researcher turns towards on developing HCI which communicates with the computer in the native language.

1.2. Motivation

The usages of speaker recognition innovation are a great extent shifted and progressively developing for the past years. Several researchers in artificial intelligence look to the organization of the brain as a model for building intelligent machines. Speech is a most natural and efficient way to exchange information among human beings. To make a real intelligent computer, it is important that the machine can perceive sound, recognize, interpret and act upon spoken information, and also speak to complete the information exchange. Therefore, speaker recognition is vital for a computer to attain the goal of natural human computer communication [16].

Researcher initiated by the relationship between MFCC algorithm and neural network model which makes the computer seems higher copy of the human mind. The relationship between MFCC based on human hearing perception and the neural network model is based on human brain also inspiring me. In the first place, the main door for the title was originated from the course HCI; instructor has been shown in class an example of speech recognition. Then after I read some related works of speech recognition finally I got speaker recognition article [9] [34] this was really motivating.

These are then causes that motivate the researcher to do in the artificial field to conduct the research speaker recognition for the local language Tigrigna.

1.3. Statement of the problem

Speaker recognition as part of biologically inspired methods of computing to be determined the next major advancement in the computing area. It highly concerns about security issue as the primary target in these sectors. Most of the researchers have found that using biometric as the security system is the most efficient method and solution of user identification and user verification through voice. Because biometric can only verify the unique characteristic, can't be a copy or steal by anyone and always with you. Speaker recognition is one of the widest research areas where biometrics techniques can be used voice for security purpose as the identification [13].

In addition to the above-stated problem, as we know Ethiopia as part of East Africa exposed to terrorist groups and sharing long side border with neighbor countries. So, this makes free zone and to defend our society from criminal acts, it needs to have a strong forensic investigation

system since some individuals may link with those groups. The proposed system is based on the premise that a person's speech exhibits characteristics that are unique to the speaker. Judges, lawyers, detectives, and law enforcement agencies have wanted to use forensic voice authentication can help the investigation a suspect speaker's to convict a criminal or discharge an innocent in court [43]. Therefore, it is crucial to develop fraud-proof identification and recognition systems to increase safety and security.

On another hand, it needs to expand of applicability of speaker identification such as remote attendance logging in selected area such as office and class, to control the flow of stored private data, confidential data and financial transactions, introducing speaker identification technology for medical remote controlling, assisting and accessing systems; allow automated control of services by voice. This can be much more convenient than traditional means of authentication which require carrying a key with you or remembering a PIN [37].

In the other case, in general researchers concern on trials to enhance the accuracy and precession of the developed intelligent system techniques because of the accuracy of the system is a major concern indeed. So, it is vital to find out the root cause of the term of accuracy performance gap for the model, recognition result mismatch problem and accuracy of identification rate are great issues still a big challenge for researchers. Finding the root cause enables to improve the accuracy of speaker identification on the previous studies [1] [7] [9].

Therefore, research aimed to address the above-stated problems and experimental development of text-dependent speaker recognition system for Tigrigna language utterance in order to find solutions to the following research questions and met the specific objects.

1.4. Research Questions

Following are the research questions, which are attempted under this research

I. How to identify the proposed system that recognizes individual speakers from a group of speakers for the same utterance?

II. Can preprocessing using EPD and noise removal techniques of the voice signal improve the performance of speaker identification rate?

III. What could be the factors variability in the performance of the speaker model result?

IV. Is ANN speaker model providing a promising result for distinguishing the identity of a person?

1.5. The Objectives of the Study

1.5.1. General Objective

The general objective of this research work is to develop a prototype for text-dependent offline and real-time speaker recognition using the artificial neural network for Tigrigna language utterance.

1.5.2. Specific Objectives

To accomplish the above aforementioned general objective, the following specific objectives are targeted:

- Review a literature, available algorithms and techniques in speech recognition in generally speaker recognition specifically ways of developing for Tigrigna language utterance.
- Collect and prepare data sets (corpus) voices recording from different peoples with different age and gender for training and testing purpose.
- Analyze the method of unvoiced speech in voice signal at the stage of signal pre-processing using endpoint detection.
- Develop a prototype for text-dependent speaker recognition system by combination of MFCC and neural networks.
- Evaluate the performance of the speaker recognition model.
- Draw conclusions from the experimental results
- Forwarding recommendations and suggestions for further research

1.6. Methodologies and Tools

Research methodologies are an essential part of any thesis. When the researcher conducted and followed scientific methods in every procedure systematically starting from data corpus collection which means sound samples data sets up to the evaluations of the prototype in order to meet the objectives of the research. The researcher also conducted tools and techniques to build the process of feature extraction and feature matching. Hereunder specify activities that are executed in the process of the research study.

1.6.1. Literature review

To understand and to gain clear representation of the work of the speaker recognition of Tigrigna language, research works in the related area reviewed. The main source of literature review includes a book, journal articles, thesis, and resource form internet and conference papers from national and international repositories are deeply reviewed. A literature review is made in the area of speaker recognition, speech recognition methods, and performance evaluation techniques. The necessary steps of speaker recognition are like the preprocessing, feature extraction and feature matching adequately reviewed.

1.6.2. Development techniques

The first step is conducting and collecting of dataset (corpus) for speaker recognition through (PC microphone and Praat sound recorder software), and capturing by considering from different ages and genders of native Tigrigna speakers. Then after proceeding to digital signal processing stage (feature extraction) and fed to the pattern matching using the artificial neural networks. In addition to the researcher used to sketch the figures applied Microsoft office Visio 2003.

1.6.3. Implementation tools

Tools and techniques for Development of prototype:

All the digital signal processing and neural network implementations carried out using the MATLAB signal processing and neural network toolboxes respectively.

Tools for analysis of feature extraction:

Mel-Frequency Cepstrum Coefficients (MFCC)

Model pattern matching used for the study:

For the proposed research modeling technique used Artificial Neural Network (ANN), model. The model designed for classification and pattern matching is multilayer Perceptron [MLP consist three layers which is, one input layer, one hidden layer, and one output layer] feed-forward neural network using back propagation learning algorithm.

Tool for evaluation method:

Evaluation of the prototype performance of recognition accuracy based on confusion matrix and commonest testing parameter evaluated speaker recognition is the accuracy of recognition.

The number of speakers testing in percentage for a given data set in classification, the accuracy of the classifier is computed as a percentage of the number of correctly classified objects with respect to the whole objects in test data set.

1.7. Scope and Limitation of the Study

The scope of this research analysis is enclosed in relation to the language of interest, type of text to be used, and methodology would follow. The research covers mainly speaker recognition with small Tigrigna vocabulary words, isolated with the phrase and discontinuous speech based prototype text-dependent speaker recognition. As the number of words that need to be recognized increases for identifying, the amount of confusion and percentage of error can also increase. The research also incorporates use one isolated word and one phrase speech is inputted for recognition and discontinuous speech could be a phrase that has words separated by silence. These characteristics more help the result effective with fewer attempts at recognition. The dataset also for both training dataset and testing dataset used consisted of speech data recorded using 40 Tigrigna words and phrases from 10 people native speaker with different at ages and genders. Testing the model includes offline, real-time and from a phone conversation. This work only considers speaker identification. Speaker verification is not included.

Hence, the research prototype studies only for the Tigrigna language not preferable to other language utterance. Limitation of national wide well prepared speaker's corpus words vocabulary since the absence of huge ready-made Tigrigna corpus for speaker recognition. The other limitation couldn't be dealing linguistics for Tigrigna language. The time limitation is obvious. Text-dependent recognition limited with shorter enrollment voices couldn't incorporate large vocabulary sets. Generally, the following are some limitation for the research.

- A personal computer with built-in microphone is used for data collection using praat voice recorder software. Other high-quality recorders were not considered during recording and real-time testing.
- It doesn't support other media files format except for .wav format.
- The speaker's health status is not considered
- Not liable to attempt mimicry.
- The numbers of samples are limited to both word and phrase.

- Not well-known national as well as regional speaker's database sound data corpus like Internationals sound databases NISP, LDC, and TIMIT.

1.8. The Significance of the Study

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.

Different individuals also take advantage of the study in order to perform secured online business transactions securely access control by voice. Hospitals, health centers, and other health service providers can benefit from the study for medical remote monitoring. Since biometrics is used as a form of identity access management and control.

Forensic investigation experts, courts, police commissions can also benefit from the study for identifying individuals in criminal justice and terrorist actions. Highly secured areas can also benefit from the study by using physical security [44]. In general, the following list will be summery of the significant of the study.

- 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.
- For criminal and surveillance investigation detection using a forensic method by analyzing and involving recorded voice samples as a proof in different trials.
- Enable users remotely communication, typing for a password without the need of input device like a keyboard, and remotely attendance methods to check the arrival and leaving times of their employees.
- Enable machines unnaturally communication with human
- Customizing services or information to individuals by voice, indexing or labeling speakers in recorded conversations or dialogues.

- 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 language Tigrigna.

1.9. Organization of the thesis

This thesis is mainly organized into 5 chapters. In the first chapter was introduced the contents of the basic introduction of speaker recognition by stated the main body the basic problem discusses the background of various concepts used in speaker recognition, and its necessity, listing of objectives with its scopes and limitations, as well as the significance of the study.

Chapter 2 gives about literature review and related works about speaker recognition. Read and citation emphasize the methodology are discussed they were followed by the researchers and journal experts. It summarizes a thorough literature review of text-dependent speaker recognition system based on the current state of the voice recognition technology. It introduces again the basic model of text-dependent speaker recognition system and its components as a means of explaining the process being carried out in sequential steps. Simultaneously it gives a complete look carefully at techniques used, work was done by other researchers, and the results obtained.

Chapter 3 discusses the methodology and strategy that the research followed to be solved by the research problem. The most applicable and different approach methods feature extractions and design for the study. In addition to this also discussed the development of a prototype, techniques of sampling the corpus (sound samples) focused in this chapter.

Chapter 4 in this chapter contains the experiment results in discussion and briefly present, analysis performance evaluation and interpretations of the data based on qualitative and proofed the concepts using MATLAB software. It gives a complete description of the proposed text-dependent speaker recognition system. It provides an in-depth look at various technical details used in evaluating the proposed model and compares the experimental results with existing work.

Finally, Chapter 5 the last chapter covered based on the chapter 4 summary of finding; made appropriate and basic conclusions out of the results of the investigation and forwarded recommendations for readers and developers in the area of speaker recognition. It concludes the research with a summary and possible future work in the field of text-dependent speaker recognition model.

CHAPTER TWO

2. LITERATURE REVIEW AND RELATED WORKS

2.1. Introduction

In this chapter presented begins by deals with the review of relevant documents including books, articles, different journals and describing the fundamentals terminology and basic concepts of speaker recognition as well as included the concept of Tigrigna language phonetics and writing system. In addition to this researcher reviewed the speaker recognition and made a comparison of the different studied an extensive literature review carried out for this thesis to dig up deeper understanding identifying and analyze in the area of speaker recognition. Speaker recognition is a generic term used for any procedure which involves knowledge of the identity of a person based on his/her voice. In this section covered some essentials and fundamentals of elements of speaker recognition, the process of sound production, basic components of speaker recognition and finally, related works they have been done by another researcher on speaker recognition are included in this chapter.

2.2. The Tigrigna Language

Language is one of conveying the information tools for humans to pass their message to others partners. Each language has got its own set of phonemes from which sounds of words are generated. Tigrigna (ትግርኛ) is an Afro-Asiatic language of the Ethiopian Semitic branch Semitic (derived from the Biblical "Shem") languages belong to a language family that includes modern languages such as Tigrinya, Tigre, Amharic Hebrew, Maltese (from Malta) and Arabic. Tigrinya is the fourth most spoken language in Ethiopia after Amharic, Oromo, Somali and It is mainly spoken in northern Ethiopia (Tigray) and Eritrea in the Horn of Africa, with around 7 million (4.5 million Tigray and 2.5 million from Eritrean total speakers (2006 – 2007 census) and the official language of the Tigray regional state.

Tigrigna speakers in Ethiopia (known as Tigrayans; Tigrawot, feminine Tigrāweyti, male Tigraway, plural Tegaru) and the Tigrinya speakers in Eritrea (Tigrinyas) and one of the official languages of Eritrea along with Arabic, There are also over 10,000 Beta Israel speakers of Tigrinya According to Ethnologue [15]. It is also spoken by large immigrant communities

around the world, in countries including Sudan, Saudi Arabia, Israel, Germany, Italy, Sweden, the United Kingdom, Canada and the United States. In Australia, Tigrinya is one of the languages broadcast on public radio via the multicultural Special Broadcasting Service [18].

Tigrinya dialects are typically split into Northern and Southern sections, which differ phonetically, accent, pronounce, lexically, and grammatically [15]. Northern dialects are Eritrean, central and East Tigray, some part of Agame and the southern dialects are Enderta Province, Tembien, Raya, Kilte Awulaelo, and some areas of Agame.

2.2.1. Phonology

Phonology deals with the assessment of the significance of phonetics and their inter-relation in a particular language or across different languages. Tigrigna language has its own characterizing phonetic, phonological and morphological properties [6]. It has a private sound owner like [?] (ዐ), [x'] (ቐ), [h] (ሐ) and son on the sound different phone from another language. Tigrigna also has its own inventory of speech sounds. For example the classification of Tigrigna consonants and vowels languages differ in their set of phonemes For the representation of Tigrigna sounds, this article [14] uses a modification of a system that is common (though not universal) among linguists who work on Ethiopian Semitic languages, but it differs somewhat from the conventions of the International Phonetic Alphabet (IPA) [6].

2.2.2. Consonant phonemes

Tigrigna has a fairly typical set of phonemes (semantically significant sounds that occur in a language) for an Ethiopian Semitic language. There is a set of ejective consonants twenty-nine and the usual seven-vowel system without considering the labiovelars ጕ [gw], ኍ [kw], ቖ [kʷ] and the fricatives ኽ[X] and ቘ [Xʷ] Thus, the maximum number of consonants would be 37 for Tigrigna. Even though there are controversies on the number of consonants, this does not affect the work of speaker recognition study.

Tigrigna consonant created airflow directly restricted that air cannot escape without creating friction that can be heard [6]. Unlike many of the modern Ethiopian Semitic languages, Tigrinya has preserved the two pharyngeal consonants. Along with [x'] (ቐ), a velar or uvular ejective fricative ኽ[X], these phonological elements make it easy to distinguish spoken Tigrinya from closely related languages such as Amharic. The appendix A shows the phonemes of Tigrinya.

However, it is harder to tell from Tigre, as this language has also maintained the pharyngeal consonants.

The Table 2-1 below shows the phonemes of Tigrinya. The sounds are shown using the same system for representing the sounds as in the rest of the article. When the IPA symbol is different, it is indicated in square brackets. The consonant /v/ appears in parentheses because it occurs only in recent borrowings from European languages. The fricative sounds [x], [x^w], [x’] and [x^w’] occur as allophones [15].

Table 2-1 Consonants with their features [6, 15]

Constant								
		Bilabial/ Labiodentals	**Dental**	**Palato-alveolar/ Palatal**	**Velar**		**Pharyngeal**	**Glottal**
					Plain	**Labialized**		
Nasal		m(ⷬ)	n(ⷩ)	ñ[ɲ](ⷭ)				
Stop and affricate	**Voiceless**	p(ⷮ)	t(ⷯ)	č [t](ⷶ)	k[ħ]	[k^w](ħ)		’ [ʔ]ⷦ
	Voiced	b(ⷧ)	d(ⷫ)	ǧ [dʒ](ⷷ)	g(ⷰ)	G^w(ⷱ)		
	ejective	p' [p’](ⷨ)	t' t’](ⷪ)	č' [tʃ’](ⷻ)	[k](ⷴ)	[k^w](ⷾ)		
Fricative	**Voiceless**	f(ⷢ)	s(ⷤ)	š [ʃ](ⷵ)	(x)(ⷸ)	[x^w](ⷹ)	ḥ [ħ](ⷣ)	h(ⷠ)
	Voiced	(v)(ⷥ)	z(ⷲ)	ž [ʒ][ⷺ]			‘ [ʕ]ⷡ	
	ejective		[s’](ⷳ)		(x') (ⷼ)	(x^w') (ⷽ)		
Approximant			l(ⷦ)	y [j](ⷿ)		w(ⷤ)		
Rhotic			r(ⷷ)					

2.2.3. Vowel phonemes

In Tigrigna, there are seven vowels. Those are generated by lung will vibrate the vocal cord since the vocal cord is slightly closed. These are ኧ, ኡ, ኢ, ኣ, ኤ, እ and ኦ. All are voiced and oral sounds. These vowels can be found in each letter, that is, each letter in Tigrigna is not a single sound rather they are a combination of two sounds, one from vowel and one from consonant [6].

These vowels can be divided into different categories depending on how they are formulated by the tongue movement and positioning occurred front, central or back position of the tongue, wideness or roundness of the constriction position, and place of the tongue (high, mid or low) [19] the lips are also a contribution when its round. Tigrigna vowels and their categorization are summarized in table 2-2.

Table 2-2 Vowels with their features [6, 15]

Vowels			
	Front	**Central**	**Back**
Close	i[ኢ]	ə [ɨ][እ]	u[ኡ]
Mid	e[ኤ]	ä [ɐ][ኣ]	o[ኦ]
Open		a[ኣ]	

2.2.4. Gemination

Gemination sometime called consonant elongation the doubling of a consonantal sound, is phonemic in Tigrinya influences the significance of words. While gemination plays an important role in the morphology of the Tigrinya verb, it is normally accompanied by other marks. But there are a small number of pairs of words which are only differentiable from each other by Gemination, e.g. /k’ɐrrɐbɐ/ (ቐረባ), ('he brought forth'); /k’ɐrɐbɐ/ (ቐረበ), ('he came closer'). All the consonants, with the exception of the pharyngeal and glottal, can be geminated [15].

2.2.5. Writing System

Tigrinya is written in the Ge'ez script (Ethiopic script), which was originally developed for the Ge'ez language. The Ge'ez script is an abugida and Tigrigna script (orthographic representation) consists of 40 core characters as shown in Appendix A. Each consists of seven vowels of Tigrinya (and Ge'ez) [see table 2-2]. Each symbol represents a consonant plus vowel syllable

[see example table 2-3], and the symbols are organized in groups of similar symbols on the basis of both the consonant and the vowel. See Appendix A are assigned to the seven vowels of Tigrinya (and Ge'ez); they appear in the traditional order. The rows are assigned to the consonants [6] [15].

This representation of each character in 7 different forms makes the number of core characters to be 244, also called syllographs (ፊደል) including twenty numbers and eight punctuation marks will be total 272 [6]. However, the pharyngeal and glottal consonants of Tigrinya (and other Ethiopian Semitic languages) cannot be followed by this vowel; the symbols in the first column in the rows for those consonants are pronounced with the vowel /a/, exactly as in the fourth row [15]. These redundant symbols are falling into disuse in Tigrinya and are shown with a dark gray background in the table. When it is necessary to represent a consonant with no following vowel, the *consonant+ə* form is used (the symbol in the sixth column). For example, the word ምስባር, *məsbar* 'to break'.

The entire table where these consonants and their orders are listed is known as FIDEL. Example Tigrigna consonants liked with vowels with their seven forms and orthographic symbols are given below:

Table 2-3 Tigrigna Orthographic Symbols

Orders	1st	2nd	3rd	4th	5th	6th	7th
Orthographic symbol	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
CV representation	ሀአ	ሁኡ	ሂኢ	ሃኣ	ሄኤ	ህእ	ሆኦ
Transcription	Ha	hu	hi	ha	he	hie	ho

Below the table is Tigrigna numbers 1-10 which is written in Tigrigna language and its transcription.

Table 2-4 Numbers from 1-10 in Tigrigna

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
hadde	kelete	seleste	Arbaate	hamushte	shedushte	shewate	shemonte	teshate	aserte
ሓደ	ክልተ	ሰለስተ	ኣርባዕተ	ሓሙሽተ	ሸዱሽተ	ሸውዓተ	ሸሞንተ	ትሽዓተ	ዓሰርተ

2.3. The Human Speech production System

The speech signal will be used for the speaker recognition task, it is very important to understand properly how voice is produced and perceived by a human being.

To achieve an understanding of human speech production, first, we should study the anatomy of the vocal system. Speech production is a multi-step process by which thoughts are generated into spoken utterances. The human voice consists of sound made by a human being using the vocal folds for talking, singing, laughing, crying, screaming, etc. The human voice frequency is specifically a part of human sound production in which the vocal folds are the primary sound source. The mechanical parts of speech production and perception are divided in two: namely the vocal system and the auditory system [12].

2.3.1. The human vocal system

The mechanism for generating the human voice can be subdivided into three parts; the lungs, the vocal folds within the larynx, and the articulators.

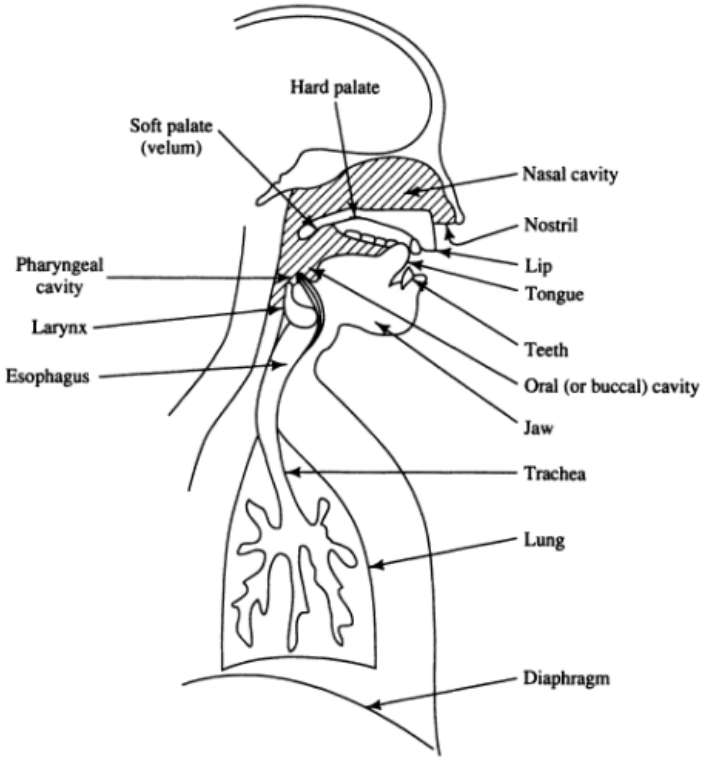


Figure 2-1 the human speech production system [12]

2.3.1.1. The Lung

The lungs are connected to trachea by two bronchial tubes which join at the base of the trachea. The lungs are used for the vital function of inhalation and exhalation of air. In the speech production model, they are the power source that supplies energy to the rest of the blocks in the systems. This airflow produced by the lungs has the shape of white Gaussian noise. The only speaker-dependent information that is introduced by the lungs is the energy of this noise. However, this is not discriminative enough and other features have to be found [12].

2.3.1.2. Vocal folds / Vocal cords and Larynx

The vocal tract is a generic term that refers to the voice production organs above the larynx. These organs are the pharyngeal, oral and nasal cavities and the velum. The length and shape of the vocal tract are highly speaker-dependent. However, their contribution to the speech signal spectrum is very important and relatively easy to measure. The vocal tract is generally considered as the speech production organ above the vocal folds consisting of [45] [10]:

- laryngeal pharynx (beneath the epiglottis)
- oral pharynx (behind the tongue between the epiglottis and the velum),
- oral cavity (forward of the velum bounded by the palate, tongue, and palate),
- nasal pharynx (above the velum just the rear end of the nasal cavity), and;
- nasal cavity (above the palate and extending from the pharynx to the nostrils).

2.3.1.3. The Articulators

Speech organs or articulators, produce the sounds of language. Organs used for speech include the lips, teeth, alveolar ridge, hard palate, velum (soft palate), uvula, glottis and various parts of the tongue. They can be divided into two types: passive articulators and active articulators. Active articulators move relative to passive articulators, which remain still, to produce various speech sounds, in particular, manners of articulation. The upper lip, teeth, alveolar ridge, hard palate, soft palate, uvula, and pharynx wall are passive articulators. The most important active articulator is the tongue as it is involved in the production of the majority of sounds. The lower lip is another active articulator. But glottis is not active articulator because it is only a space between vocal folds [17] [38].

2.3.2. The human auditory system

The complete auditory system includes the whole ear assembly, auditory nerve bundle and the auditory cortex of the brain. These different parts of the human auditory system are responsible for registering the audio signal in our brain. Then after, the brain which is responsible for interpreting these signals into parts of the language and speaker identity [13].

2.4. Neural Networks

A human brain works differently than a typical computer machine. While a computer can use a processor to execute instructions and access local memory at incredible speeds, a human brain is a slower collection of nodes called neurons. These neurons are connected by synapses which pass electronic or chemical signals back and forth to each other. These connections are capable of changing each time information is sent or received which represents the brain adapting and learning information. A neural network is an artificial computer generated system that attempts to replicate the neural system of the human brain. Nodes are used to represent neurons, and they are connected with different weights to represent the synapses. These networks can be trained to process different information and learn from it. Naturally, these systems are used to mimic certain functions that the human brain is capable of, such as pattern recognition and decision making [5].

2.4.1. Biological Neural networks

A biological neural network is a series of interconnected neurons whose activation defines a recognizable linear pathway. The different brain functions emerge as a result of the dynamic operation of millions and millions of interconnected cells called neurons. Roughly speaking, a typical neuron receives its inputs through thousands of dendrites, processes them in its soma, transmit the obtained output through its axon, and uses synaptic connections to carry the signal of the axon to thousands of other dendrites that belong to other neurons [17] [39].

It is the manner in which this massively interconnected network is structured and the mechanisms by which neurons produce their individual outputs that determine most of the intelligent behavior. These two aspects are determined in living beings by evolutionary mechanisms and learning processes [17] [39].

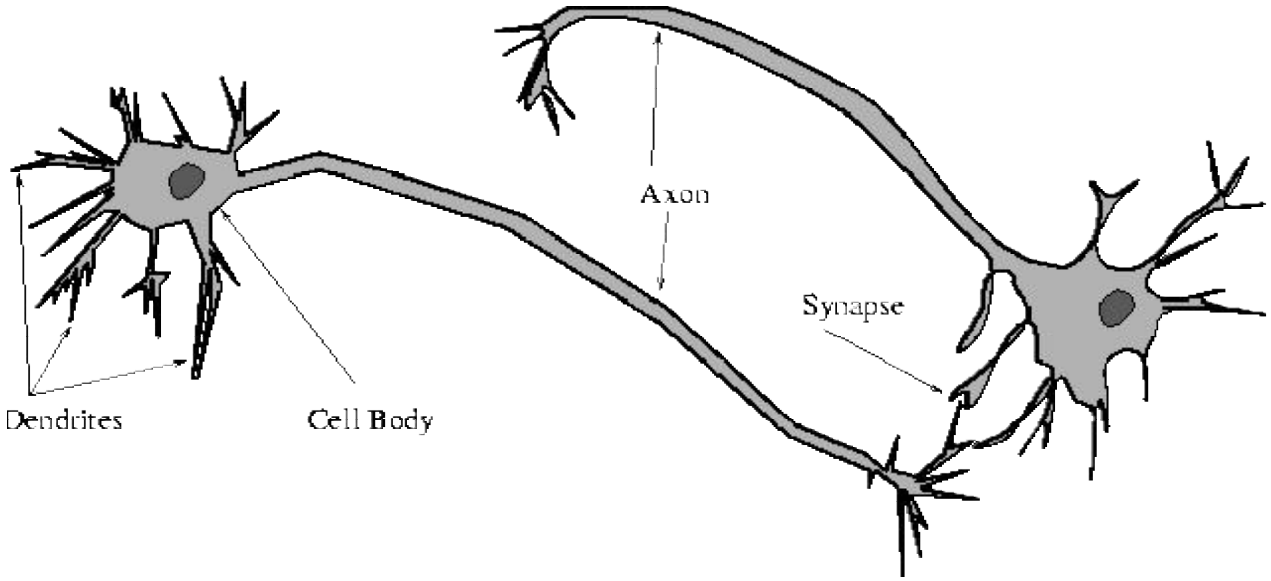


Figure 2-2 Simplified Biological Neurons [17]

2.4.2. Artificial neural networks

The structure of artificial neural networks modeled based on the basic property and understanding of biological neural systems. The central idea behind most of the developments in neural sciences is that a great deal of all intelligent behavior is a consequence of the brain activity [17].

The model of the entire artificial neural network is determined by the network topology, type of neural model, and learning rules. These are the main interests in designing artificial neural networks [41].

A typical example is Multi-Layered Perceptrons (MLPs) are one of the simplest types of neural networks, comprised of several neurons arranged in different layers (see Figure 2-3 below). The layer that consists of the neurons which receive the external inputs is called the input layer. The layer that produces the outputs is called the output layer. All the layers between the input and the output layers are called hidden layers. Any input vector fed into the network is propagated from the input layer, passing through the hidden layers, towards the output layer. This behavior originated the other name for this class of neural networks, namely. Feed-forward neural networks [17] [46].

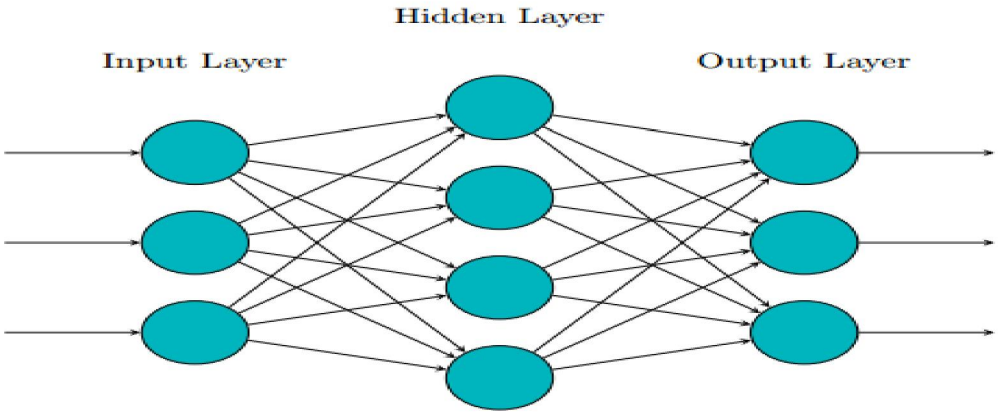


Figure 2-3 MLP feed-forward artificial neural network with one hidden layer [17]

2.5. Basic Components of speaker recognition

Generally, speaker recognition is the method of automatically identifying who is speaking based on his individual speech wave information [20]. The crucial aim of speaker recognition has two phases. These are training and testing phase:

The following diagram shows the basic steps of a speaker recognition system as a general.

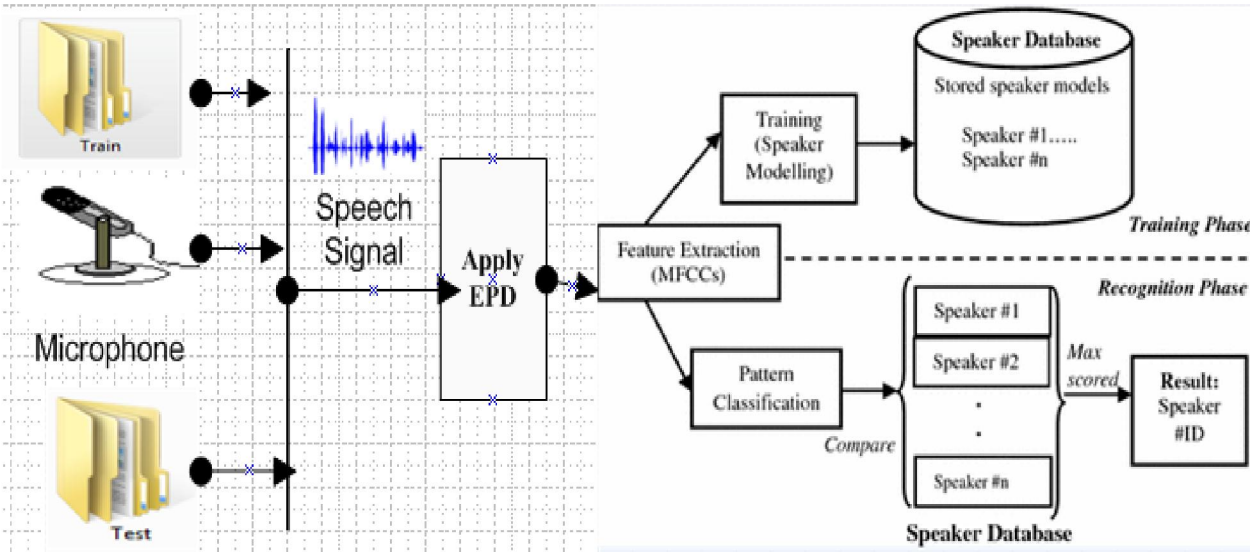


Figure 2-4 Training and Testing Phases for speaker recognition [20]

I. Enrollment session or Training phase

In the first place during training phase or at enrollment session models, speakers' information voice print is created for each speaker by extracting features from the speech signal stored in the sound database. Then, the system can build a reference model for those individual speakers.

During the training phase, the voice information is detected and compared with the information stored making the system know the speakers and deals with collecting data from the utterances of people to be identified.

II. Operation session or testing phase

During the testing phase, the unknown claimant features or utterance are compared against with the trained speaker's features means previously created or stored in the database voice print and then recognition is made from the models. Simply it's a task of identifying an unknown utterance from groups of known speakers.

For identification systems, the utterance is compared against multiple voice prints in order to determine the best match likelihood while verification systems compare an utterance against a single voice print to give a decision either a claimant or imposter [21].

2.6. Classification of speaker recognition

Although the terms speaker identification and speaker recognition are used interchangeably; speaker recognition as generally is broader than identification.

Figure 2.5 below provides the various classifications of speaker recognition. Speaker recognition systems can be classified attending to three different characteristics. Firstly, it can be divided based on major tasks into speaker identification or speaker verification. Secondly, it can be divided based on methods of text uttered into text-dependent or text independent systems. Finally, it is possible based on the set of trained speakers classify them into open-set or close-set systems.

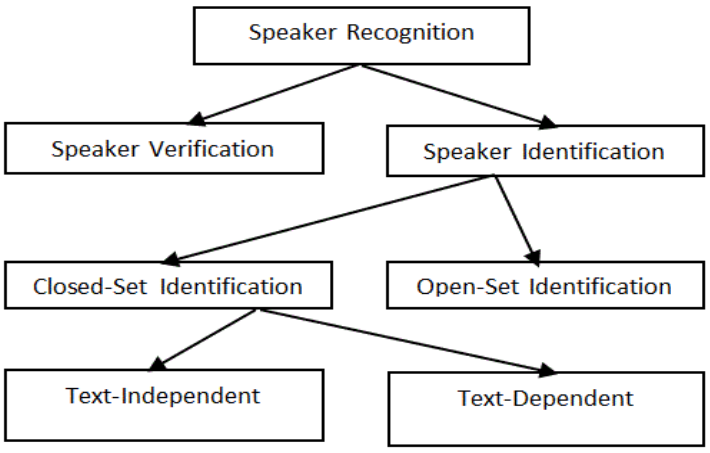


Figure 2-5 Speaker recognition classifications [22]

I. Identification vs. Verification

The task of speaker recognition can be broadly categorized into two types:-

1. **Speaker identification:** It is the process of multi-class classification problem which is registered speaker provides a given unknown test utterance assigned to a particular class. In the speaker identification system the test utterance is compared the inputted one speech uttered with all registered sound speakers and scored with the registered speakers and the one those modules matches the best is selected. This is done by comparing the pattern of the speaker's voice with a trained and stored speaker database. The database contains reference models of all known speakers voice acoustic vector [25].

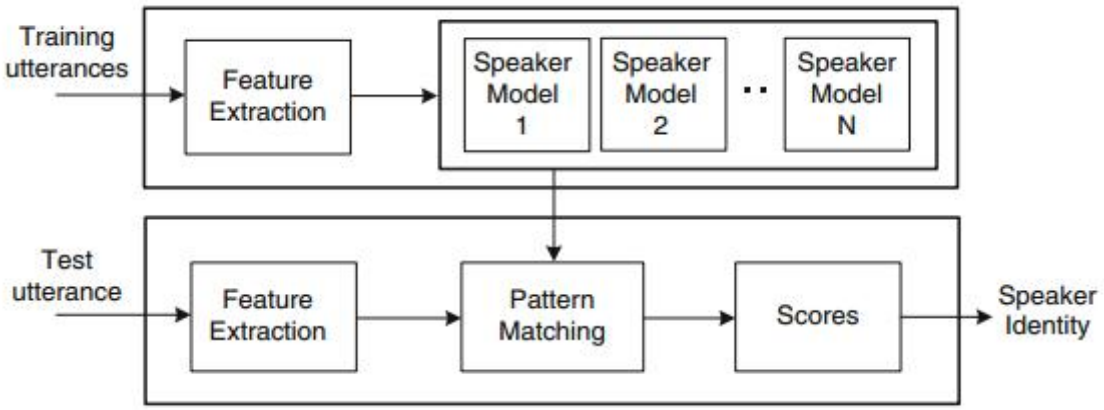


Figure 2-6 Speaker Identification Diagram [42]

2. **Speaker verification:**

Speaker verification is the process of accepting as claimant or rejecting as imposter the identity claim of a speaker. It is a binary classification problem and one to one comparison then after a single decision based on the result of that comparison. The system extracts parameters from the input speech signal to represent vocal characteristics and uses this information to build representative speaker models. 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 assumed imposter the test utterance, otherwise the system will accept the test utterance claimant. One of the major challenges in speaker verification is trying to find a reliable threshold that can be used for decision making [25] [22].

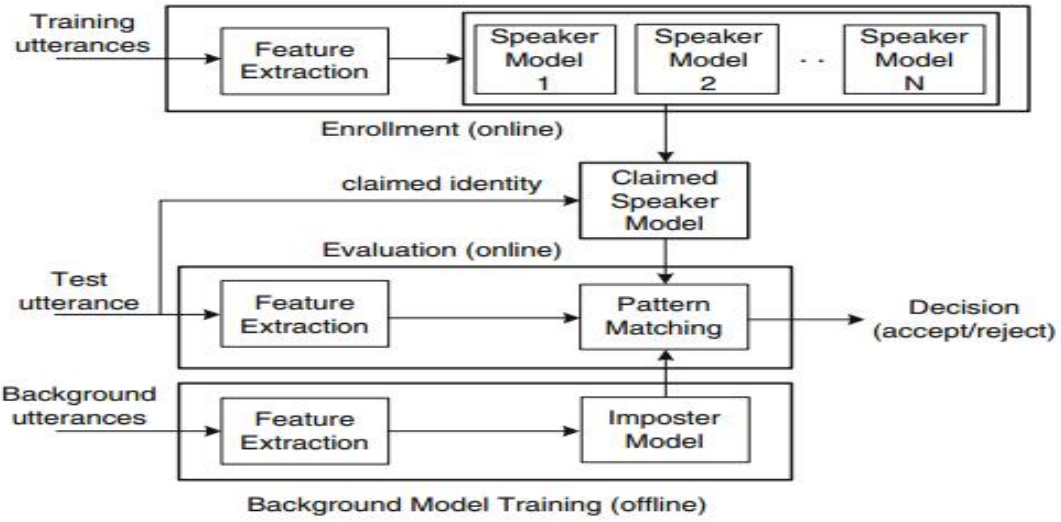


Figure 2-7 Speaker Verification Diagram [42]

II. Text-dependent vs. Text-independent

This is another category of classification based on the method of text uttered or constrained sentence which is used for speaker recognition systems. This category is based upon the text uttered by the speaker during the identification process.

1. **Text-Dependent:** In this case, the test utterance is identical to the text used in the training phase. The test speaker has prior knowledge of the system and user cooperative. Text-dependent requires the speakers to say exactly stored in the database or the same speech. In this process, the speaker is asked to utter a prompted or a fixed phrase. This type of recognition

has an advantage of increasing the performance of the system because of the prior knowledge of the spoken text. The larger number of keywords used, the more robust system can be built [23].

2. **Text-Independent:** In this case, the test speaker doesn't have any prior knowledge about the contents of the training phase can speak anything and users are uncooperative. Text-independent does not require and does not care the speakers to utter the same speech and the message content; the speaker can be recognized by attending to speaker-dependent attributes of speech. Here the system does not know what the speaker will say; a more flexible system will be needed [23].

III. Open set vs. Closed set

Speaker recognition can be classified into open set and closed set speaker recognition based on the set of trained speakers available in the system.

1. **Open Set:** An open set system can have any number of trained speakers. The system should be able to determine whether a speaker is enrolled or not (impostor) and, if enrolled the system will determine his or her identity. In this process, the system does not know if the test utterance belongs to the registered speakers or not. So if the system does not find any register speaker who is matched with the test, the system will reject to assign the utterance to any of the speakers [25].
2. **Closed Set:** A closed set system has only a specified (fixed) number of users registered to the system by training the network and stored. Identifies the speaker as one of those enrolled, even if he or she is not actually enrolled in the system. In this case, closed set there is no rejection scheme at all [25].

2.7. Modules of Speaker recognition

The process of Speaker recognition consists of 3 modules namely: - preprocessing, feature extraction and feature matching.

I. Pre-processing

Pre-processing is the first step for signal processing technique of speech. The process of preprocessing in speech recognition discussed in the study includes Noise removal using spectral subtraction, Pre-emphasis, and endpoint detection. It is considered the first phase of other for the

next feature extraction phases in speech and speaker recognition to differentiate the voiced or unvoiced signal and filtering finally pass to feature extraction.

II. Feature Extraction

Feature extraction is the process in which we extract a small amount of data from the voice signal that can be used to represent each speaker. The purpose of this module is to convert the speech waveform into a set of features or rather feature vectors used for further analysis. As its known speech signal is a called a quasi-stationary signal i.e. a slowly timed varying signal. It needs some analysis on the signal to get the characteristics information carried when examined over a short period of time (e.g. between 20 and 30 ms).

Therefore, the short-time spectral analysis is the most common way to characterize a speech signal [11]. A wide range of possible methods exists for parametrically representing the speech signal for the speaker recognition purpose. These include Linear Prediction Cepstrum Coding (LPCC), MelFrequency Cepstrum Coefficients (MFCC), Perceptual Linear Predictive (PLP) and Wavelet [25].

III. Speaker modeling and Feature matching

After the feature extraction process finished, feature matching task of the classifier to carry out speaker recognition. For this, it needs to model each of the speakers enrolled in the training phase for speaker identification and speaker verification. In the testing phase, the speaker identification system compares the sequence of feature vectors representing the test utterance with each of the speaker's model and decides the identified speaker to be the one whose model shows maximum similarity scored.

In speaker recognition, there are two basic classifier models. Template model based classifiers and stochastic model based classifiers. The most common template model based classifiers are dynamic time wrapping (DTW), vector quantization (VQ) and neural networks (NN). Examples of stochastic based model classifiers are Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) [13].

The method using neural networks as an intelligence pattern comparison technique was the best method in terms of flexibility, adaptability, and modernly. Neural networks gave a higher performance in comparison to other classification techniques in many studies and researchers.

The MLP architecture is suitable for both the features set and the selected speakers were to be designed and tested. In this research study, the concepts of both MFCC and neural networks were implemented [33].

2.8. Comparison of biometric authentications

A biometric system is a pattern recognition system which studies on physical and behavioral characteristics of human beings. This makes a personal identification by determining the authenticity of a specific physiological or behavioral characteristics possessed by the user. It comprises methods for distinctively recognizing humans based upon one or more natural physical or behavioral traits. Biometric is used as a form of identity access management and access control. It is also used to identify individuals in groups that are under surveillance [22]. Biometric characteristics can be divided into 2 main classes:-

I. **Physiological**

A physiological biometric is based on the shape of the human body. This system uses prior knowledge of the human traits for classification. Fingerprint, face, D.N.A, palm print, hand geometry and iris recognition systems are classified as physiological biometrics.

II. **Behavioral**

A behavioral-based identification performs the task by recognizing people's behavioral patterns. Rhythm, gait, signature, and voice (speaker) recognition are mainly based on the study of the way a person behavior, commonly classified as behavioral. Features that uniquely represent the characteristics of individual speakers' voices are estimated and used for classification [22].

Biometric-based authentication measures individuals' unique physical or behavioral characteristics. Among the above, the most popular among all biometric system is the speaker (voice) recognition system because of its easy implementation and economical hardware [22].

However, fingerprints and retinal scans are more reliable means of identification, speech can be seen as a non-evasive biometric data that can be collected with or without the person's knowledge or even transmitted over long distances via telephone could be remotely accessed without need of physical contact. Biometric authentication has some key advantages over knowledge and token-based authentication techniques. Unlike other forms of identification, such as a PIN, a person's voice cannot be stolen, forgotten or lost, and no need of carrying. It has to be easy to extract, measure, save and compare with less expensive hardware. Meanwhile, voice

can easily affect because it depends on health condition, quality of microphone and background noise. Speech utter has never been among people that have the same voice. If someone wants to copy and mimicry it by recording, the machine or equipment that been used must have a very high quality. Speaker recognition with proper statistical, analytical and data processing techniques thus allow for a secure method of authenticating speakers [26][24][27].

2.9. Factors affecting speaker recognition performance

Factor that should be considered for development of a robust speaker recognition system like speech recognition, the following cases are few they play a significant role in the system state of accuracy:

- The number of vocabulary with their variability (Isolated, discontinuous, or continuous speech) and speakers size
- Variability in speakers: Gender, speed, regional and social dialects and language variability within the same language, speaking style, emotional, physical states and volume of the loudness
- Variability in recording environments: background noise, reverberation (echo)
- Variability in transmissions: quality of channels and microphones
- Mismatched training and test conditions
- Noisy voice samples are difficult to analyze background noise.

2.10. Related Works

The following different researchers have conducted their studies on automatic speaker recognition to identify the identity of individuals from their voice speeches. In this section, related works performed to identify speakers on different languages are discussed.

Hafte Abera [6] implemented Hidden Markov Model Based Tigrigna Speech Recognition using Hidden Markov Model Toolkit Perl programming language. This study aimed to design speaker independent continuous Tigrigna recognition system. The speech recognizer is then evaluated using the test dataset performance result shows 60.32% word level correctness, 58.38% word accuracy, and 20 % sentence-level correctness were obtained. Hafte indicated in the recommendation there is a need to undertake research using neural networks and mixed types which contain both HMM and neural networks to improve their performance [6].

Aykefam Azene [1] developed text-independent speaker identification system for the Amharic language. For his identification process speech signal collected at a sampling frequency of 16 KHz and 16 bit. Then after MFCC, ΔMFCC, and ΔΔMFCC used for feature extraction. After these features are extracted from each speech signal, vector quantization and gaussian mixture models have been used for training and identification purpose. Aykefam used for his research total sound dataset 50 speakers for text-independent speaker recognition. Finally, he got 74.2% accuracy was achieved when vector quantization approach where used whereas 84.3% accuracy for the gaussian mixture models. At his future work raised his thesis as a start point for researchers who are interested to do their work on the area of speaker identification.

Whole law Berehane [45] designed and implemented Amharic Text-Prompted Speaker Verification Model. Author's model applied frame-based processing to the speech wave signals so that all samples extracted using MFCC. After the feature extraction made researcher used Support Vector Machine modeling for verification for each speaker. Author's dataset was collected Amharic word was uttered ten times repeatedly by 5 men and 5 women totally 100 speech wave files recorded. Performance is measured in terms of False Acceptance, False Rejection, True Acceptance and True Rejection. Finally, ten Amharic words experimented, an average performance of 0.25% EER, 99.7% accuracy, 99.8% recall and 99.7% precision is obtained and want to extend for other language followed excellent accuracy performance.

Sanaa Hamid Helmed Mohammed [33] has studied on Speaker Identification System using Wavelets and Neural Networks to provide an accurate and efficient identification. Six words were recorded by a total of 4 speakers and their speech signal frames were 256 with 88 samples overlapping and hamming window. Feature extraction was implemented with the aid of the wavelet packet transform. A Neural Network-based approach was implemented for identification and classification of the extracted features as belonging to the speakers.

The selected architecture was a feed-forward neural network with one hidden layer. To train the neural network, gradient descent back-propagation with momentum was used. The system after training gave an accuracy of 93%. For the real-time hardware implementation of the system, a Digital Signal Processor from Texas Instruments called the DSK6713 was used.

Chularat Tanprasert [9] has provided a work on neural network based text-dependent speaker identification system for the Thai language. The author used feature extraction from the speech linear prediction code that formed feature vectors. Those features fed to multilayer perceptron

neural network using back propagation learning rule for training the network. Average identification rate on 9 speakers achieves above 95% when using mixed tone text, and poor results occur with middle and low tone texts, which usually cause vagueness or unclear voices.

The researcher has been done differ from those listed above by applied MFCC feature extractions to multilayer Perceptron neural networks. By identifying the gap finding the causes and improving the performance of evaluation method and capabilities of speaker identifying ANN has a great power for nonlinear recognition performance based on pattern recognition template approach.

2.11. Summary

In this chapter presented the literature review starting by described local language Tigrigna, the formation of the speech productions, the basic components, the various classifications and modules of an automatic speaker recognition system. At last, the researcher reviewed, understanding the gap and factors which was affected the performance of the different related works.

Therefore, in this thesis, it's described some methods of speech preprocessing, feature extraction and speaker models. The researcher was chosen for the study from speaker recognition tasks the speaker identification, text-dependent among the method of text uttered, from the set of trained speaker designed a closed set, for the met of objectives, feature extraction were used MFCC and for feature matching and classification ANN model. Text-dependent and a closed set of trained speakers were selected gave output the most likelihood and closed one.

Researcher preferred MFCC is one of the widely and most popular with a robust technique for feature extraction used systems is the voice recognition technique based on perceptual frequency scale human hearing.

Many algorithms are currently used for searching the best matching path between two signals. Some of the more successful techniques include dynamic programming; hidden Markov models and neural networks applied at both the phoneme and at the word level. However, neural networks remain the most widely used algorithm for real-time recognition systems and selected by the researcher. It is considered sufficiently mature to solve sequential decision problems and to prove its functionality and the usage of neural networks for the study.

CHAPTER THREE

3. METHODOLOGY

3.1. Introduction

This chapter explained and presented methodology of the speaker recognition by stating the research design, process model, ways of data collection, and sampling with their techniques. It focused on the design issues and process model, algorithms and techniques that are used for pre-process an audio signal before it's further processed by applications of feature extraction to feed the neural networks.

The methodologies were developed for speaker recognition steps are usually divided into three phases; namely preprocessing phase, modeling phase and matching phase. Preprocessing is the stage of converting from analog signal speech to digital sample frames. The modeling is a process of enrolling speaker to the identification system by constructing a model of his/her voice, based on the features extracted from his/her speech sample. The matching is a process of computing a matching score, which is a measure of the similarity of the features extracted from the unknown speech sample and compare with a reference stored in the database speaker model.

3.2.Research Design

After spending time on reading and digesting the literature in a particular research area, it needs further investigation and identified some potentially important variables and probably has become familiar with the methods commonly used to measure that behavior. Research design means the overall strategy of that we chose to integrate different components of the research in a coherent and logical way to ensuring effectively address the research problem to constitutes of the variables.

Choosing an appropriate research design is crucially important to the success of the project. The decisions you make at this stage of the research process to determine the quality of the conclusions can draw from your research results. As a scientific study, this research exploratory data collection and analysis, this is aimed at classifying behaviors within a given area of research, identifying potentially important variables, and identifying relationships between those variables and the behaviors. Such exploration is typical of the early stages of research in the area of speaker identification rather than, hypothesis testing [29].

Finally, the researcher chosen research design for this case of study quantitative, experimental approach by exploratory data collection and analysis to address the research problem is mentioned in chapter one section 1.3.

3.3. Process Model

The process model is based on the research design which is the research activities from the initial problem specification through design, implementation, and evaluation of solutions. The process involves dataset preparation, design and implementation of a prototype and evaluation of the prototype to measure its performance. The processes of the prototype model as a summary procedure and over all of the steps for identification.

I. Audio recording using praat for training and testing save as classified into two categories, they are: for training in training folder and for testing in testing folder.
II. From microphone record a sound real-time for given duration, it has live audio speech unvoiced part of speech detection and then either save in training folder for training latter use or real-time test the network model.
III. Noise reduction can apply if necessary using spectral subtraction during recording using praat, the speech segmentation and Feature extraction using MFCC from recorded audio and stored as a reference model in Database in the form of sample frames for each speaker specified with a class ID number.
IV. Neural network, training group and target group assignment
V. It has a mechanism listening to checking the speech sound whether appropriately selected the needed audio with its information from the imported file and after-life recorded via microphone.
VI. Run neural network on new audio test folder or real-time with respect to trained set which is stored in the sound database as template reference model in the form of vector.
VII. Neural networks compute the maximum scored likelihood unknown speaker with known speakers identified and display speaker class ID with its identifier name.

As it is already mentioned in the above chapter two section 2.5, a biometric system, it has two sessions: - this speaker recognition system can work into two different modes called training phase and testing phase. During the training phase, the speech is trained to the system to build a

reference model for the speaker. In the training phase, the features of the speech are extracted using the MFCC algorithm and then stored in the database for future or later use. During the testing phase, the input speech is matched with the stored reference template model. Biometric information is detected and compared with the information stored at the time of enrollment. The input speech and the reference template model are matched based on pattern comparison using artificial neural network pattern matching algorithm.

3.4. Data Source

One of the greatest challenges in the field of speaker and speech recognition is the lack or absence of open source data (corpus) or public database readymade here in Ethiopia for national as well as local language. There have been quite a few public databases collected in the past decade internationally like NIST, LDC, and TIMIT. So, the main source of as primary data source is directly audio records utterance wave file recorded and saved separately from 10 individual people (5 male and 5 female). From each got four time at the different time record total sound corpus is 40 samples some of them are also scraping audio from conversation telephone speech. Researcher indeed promised to those who willingly record their sound could ethically use their voices.

The word selected by the researcher is one single word (ተጋዳላይ) Tegadalay (means Soldier) and one phrase (ጋዜጣ ወይን) Gazeta weyin (means a name of the newspaper is publishing in Tigray). The reason selected the word ተጋዳላይ is marvelous popular and known Tigrigna song and the phrase ጋዜጣ ወይን is also almost known newspaper by individuals weekly publishing and distributing written by Tigrigna language considering a common generalized and standard to all native for Tigrians.

Finally, the word and phrase are selected and prepared during training and testing for the speaker recognition model.

3.5. Speech Samples Collection Techniques

Speech samples collection is a very important step towards producing automatic identity recognition systems with efficient performance. In order to collect the dataset corpus voice speech, two methods were used. The sound corpus collects directly record from individual speakers and from telephone conversations. The speech samples are varied in terms of age levels stripling or young, youth and teenager. The speakers are selectively chosen from a different part

of Tigray area such as Raya, Mekele, Adwa and Shire for the matter of representation coverage and considering the variety of accents. Total 10 (5 male and 5 female) each four-time utterance person considers for this research. Each sample is taken at a sampling rate of 44.1 KHz and quantization resolution of 16 bits per sample in the stereo channel. After being collected, all these data is properly preprocessed and the necessary features are extracted.

According to the above description of the way of speech samples collection is mostly concerned with recording various speech samples of each same word by different speakers following different criteria's. The following four factors considered when collecting speech samples for training and testing set. The first, speaker profiles those speakers include belonging to different ages, genders, and area. Table 3-1 summarized speakers' profile information below.

Table 3-1 Summary of speaker profile

Range of age	sex	area
15-22	1 Female	1 Mekele
22-38	4 Male and 3 Female	2 Raya,1 Mekele 2 Adwa , 2 Shire
Above 38	1 Male and 1 Female	1 Mekele, 1AbyiAdi

Secondly, speaker condition area. This is basically refer to the environment of the recording stage. The researcher tried to record using praat software the speech from the silent room and the less noisy environment is included.

The third factor is the transducers and transmission systems. Speech samples were recorded and collected using a normal personal computer microphone and from a cell phone.

The fourth factor is vocabulary size. The speaker recognition is specific one isolated word and one isolated phrase [32].

3.6.Sampling Techniques

The researcher chose one of the many sampling techniques is judgmental sampling (sometimes called purposively sampling technique). This kind of sampling technique is preferable for this kind of research which is considered for millions of populations taking a sample. In this type of sampling, subjects are chosen to be part of the sample with a specific purpose in mind assumed they would be perfect for its sample of speech corpus. The researcher believes that some subjects are fit for the research compared to other sampling technique. This is the reason why

purposively chosen as subjects however the drawback of the technique is its subjectivity. As a summary of the why because of chosen of sampling techniques are:-

- The researcher aims to do a qualitative, experimental and exploratory study as selected in the research design above section 3.2. And this type of sampling can be used when demonstrating that a particular trait exists in the populations.
- The research also considers randomization is impossible like when the population is almost huge around from millions taking 10 persons.
- The researcher shows the usefulness has been done with a limited budget, time and workforce.

3.7. The proposed speaker recognition architecture of the System

The human voice is a complex information-bearing signal, depending on physical and behavioral characteristics. The raw speech signal, uttered by any person, is extremely rich in the sense that it involves high dimensional features. To perform efficient speaker recognition, one must reduce this complexity, while keeping sufficient information in the extracted feature vector [26].

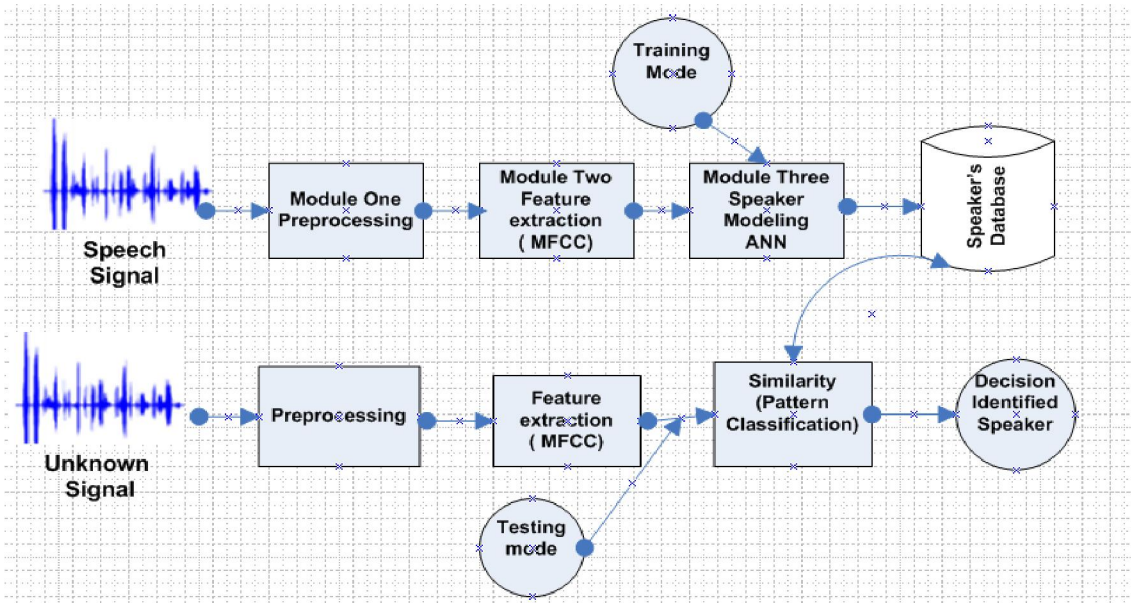


Figure 3-1 Proposed architecture of the speaker recognition

A simplified the three layer architecture of the speech processing module is presented here in Figure 3-1 is modular and the implementation of each of the subsystems shown is also modular. This architecture is showing then used for both training and testing modes. This audio signal

after preprocessing module has been done then fed to the feature extraction module. Once the feature set is extracted from the input signal, it is then sent to the Classification module, which is responsible for measuring the similarity of the extracted feature set to the models already present in the reference set. The computed measure of similarity is passed on to the decision module.

The original speech signal is not a stationary signal since the vocal tract is continuously deformed and the model parameters are time-varying. But, it is generally admitted that these parameters are constant over sufficiently small time intervals digitized by sampling it at the desired frequency. Classically, the original signal is divided into frames of 30 milliseconds denoted. Then after applied hamming windows is applied by overlapping the fill the gap of discontinuity of the signal.

Feature extraction based on the speech signal registration and preprocessing, features are extracted to define a model corresponding to the user. Ideally, these features must be robust to intrinsic variability of the user's voice, to noise and distortion, to impersonation. The most widely employed methods involve short-term spectral features. We have chosen to extract MFCC (Mel-frequency cepstral coefficients) which reveal to be more robust and efficient in practice. Once these features have been extracted stored in the speaker model each sample as a feature vector, the corresponding model or template design requires a training phase. Again, we had chosen the artificial neural networks method [47]. Decision logic - makes the final decision about the identity of the speaker by comparing unknown feature vectors to all models in the database and selecting the best matching model.

3.8.The development of the speaker recognition system prototype

The way of data collection techniques and data sources are already stated above. After the necessary features are extracted, different sequential steps were performed to achieve the goal of the thesis. The goal of this study is to build a simple to representative automatic speaker recognition system based on neural network for Tigrigna language utterance. The data sets divided during train speaker model into training 70 percent, of the total sample and the remaining 15 percent is used for testing and 15 percent for validation purpose. Appropriate training and testing algorithms were selected, training and testing were performed. Known speakers are classified into classes by specifying ID number with optional names and the speech feature of each speaker is tested against the available speaker classes to identify the identity of each

speaker. Before all these given results, each sequential step and their required elements are well stated.

3.8.1. Raw Speech and Audio file format

3.8.1.1. Raw Speech

The raw speech in speech signal wave could be in the form of time domain or frequency domain. Samples are generated from the analog signal; the audio signal captured by the microphone is sent through the classification module to identify whether the captured audio recording is silence, has speech content, or is noise. The speech signal has an enormous capacity for carrying information. Speech is a continuous signal but is stored electronically as discontinuous samples. This process of reducing a continuous signal to discrete sequence numbers is called sampling to represent the original signal.

3.8.1.2.Audio file formats

Once the speech signal has been sampled it is stored in the form of samples in a file. One of the most popular encoding formats WAVE (.wav) format. Wave file formats differ in the amount and type pure (lossless) of compression, their headers and some other specifics. The standard format for a wav file is 44,100 samples per second and 16 bits per sample. Information such as a number of channels stereo, sampling rate, data size, etc is needed to make any meaningful use of the data. Since each sample is stored in 2 bytes, each speech application first re-constructs the samples and stores them in an array. Pre-processing and further processing are then performed on arrays thus obtained.

3.8.2. Preprocessor

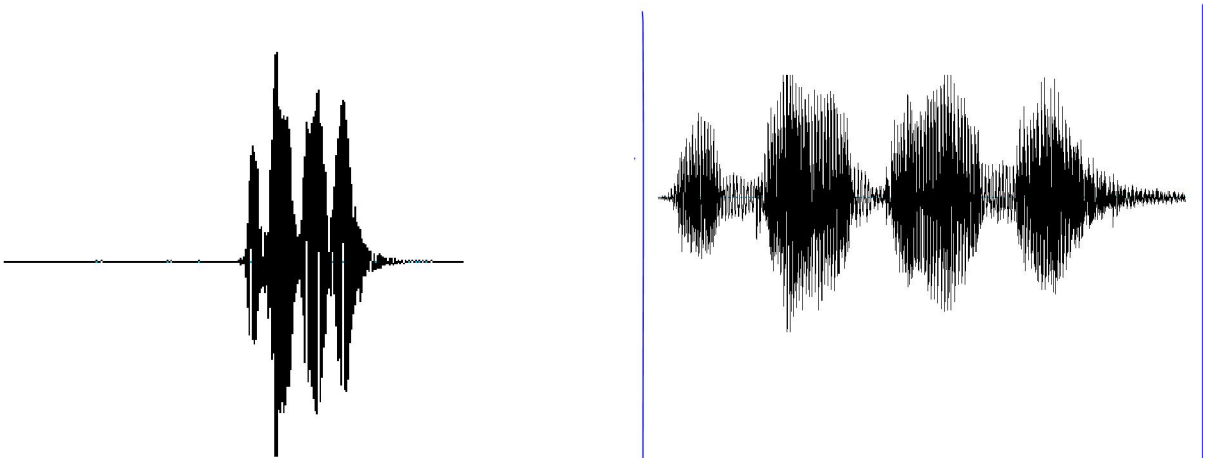
Pre-processing is performed before any speech specific information is extracted from it. In this work, we considered the preprocessing as the first step for speech signal processing. Preprocessing is generally performed initially to determine if an audio signal contains any speech content. Once speech content has been determined, emphasizing the speech content and endpoint detection may also be performed during the preprocessing stage itself.

The main purpose of preprocessing is to convert continuous time signal sampled at discrete time points to form a sample data signal representing the continuous time signal. Then, samples are quantized to produce a digital signal [32] and to reduce the amount of data to be processed by

the rest of the system. Preprocessing involves Analog to digital conversion and Pre-emphasis filtering [26].

3.8.3. Silence Removal and End Point Detection

In this stage followed from preprocessing is detecting the problem in speech processing presence of background noise from the utterance called endpoint detection. 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. The accurate detection of a word's start and end points means that subsequent the processing of the data is to keep it to a minimum number of samples. This is a fundamental step for applications like Speaker Identification. Speaker Identification also demands efficiency in feature extraction even in the noisy environment [35].



(a) (b)

Figure 3-2 Before EPD (a) and after EPD (b) for the word Tegadalay

During implementation and practically shown that endpoint detection is particularly difficult to do accurately if the speech is uttered in a noisy environment. In the case of unvoiced (noisy) sounds occurring at the beginning or end of an utterance, it is difficult to distinguish accurately the speech signal from the background noise signal.

3.8.4. Feature Extraction Module

The feature extraction process transforms input waveform into a sequent of the acoustic feature vector in which speaker specific properties are emphasized and statistical redundancies are suppressed. It considered as a data reduction process that attempts to capture the essential characteristics of the speaker with a small data rate. The importance of feature extraction is

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. After the features are selected from the utterances, a method for classifying them as belonging to the speaker must be selected [33]. The objective of modeling technique is to generate models using speaker-specific feature vector.

For speech and speaker recognition, the most commonly used acoustic features extraction are Mel-scale frequency cepstral coefficient (MFCC).A “mel” [mel come from the word melody] is a unit of special measure or scale of the perceived pitch of a tone. MFCC takes human perception sensitivity with respect to frequencies into consideration, and therefore MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. MFCC features are the most commonly used features in speaker recognition. It combines the advantages of the cepstrum analysis with a perceptual frequency scale based on critical bands [33]. MFCC consists of 6 computational steps as shown below figure 3-3.

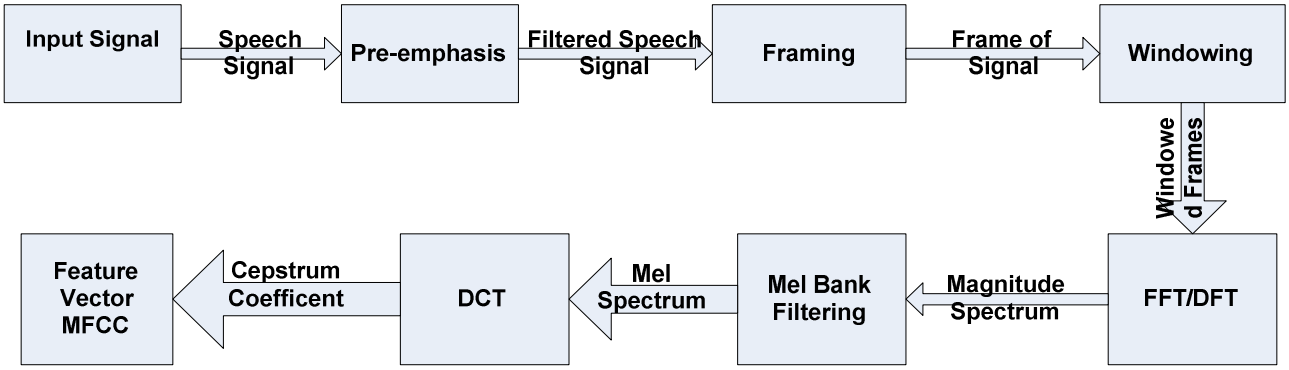


Figure 3-3 Block diagram for MFCC steps

Step 1: Pre-emphasis filtering

This step processes design to the passing of signal through a filter which emphasizes higher frequencies. This process boosts the energy signal at a higher frequency since energy is higher at the lower frequency. This is caused by the nature of glottal pulses. The idea of pre-emphasis filter is to cancel the glottal effect from the transfer function of vocal tract by applying high pass filtering. The pre-emphasis is done by using a filter was taken a coefficient of 0.95. EmphasisSignal(n) = Signal(n) - a *Signal(n -1) where is the input speech data and is the output

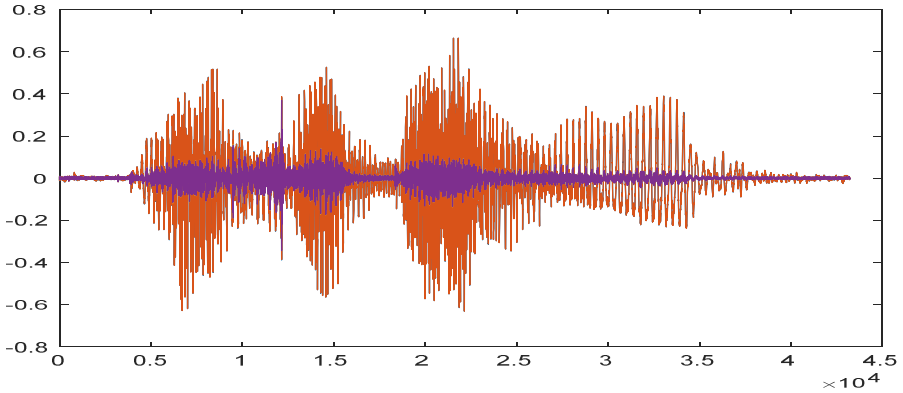


Figure 3-4 Figure before (red color) and after the Pre-emphasis for Gazeta weyin

Step 2: Framing

The speech signal is not a stationary signal since the vocal tract is continuously deformed and the model parameters are time-varying. But, it is generally admitted that these parameters are constant over sufficiently small time intervals. Classically, the signal is divided into frames of 30ms. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N).

For the frame sample of 44100 are to frames size by 30 milliseconds, are 1323 samples number of frames 19 and The input speech signal is segmented into frames of 30 ms with an optional overlap of 1/3~1/2 of the frame size.

Typical values for N and M are N = 1323 and M = 500.

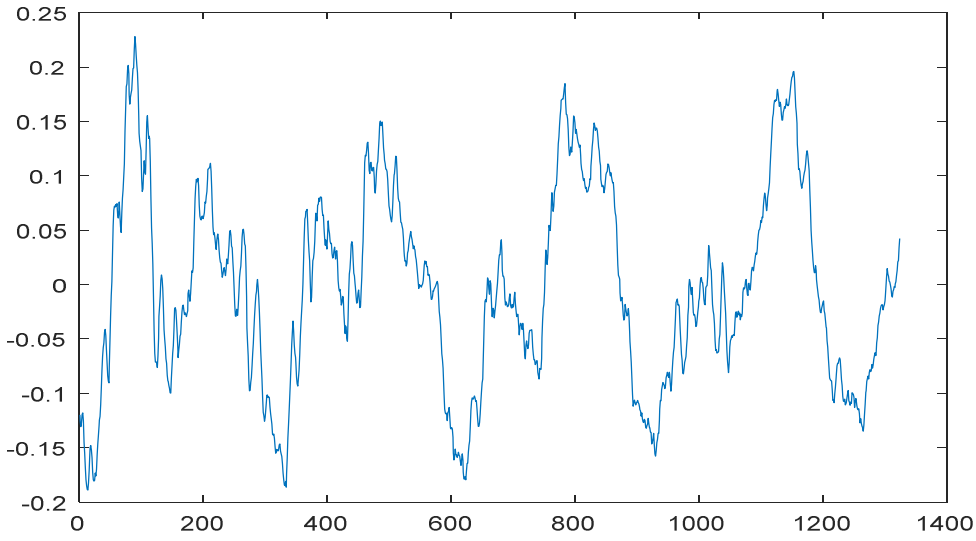


Figure 3-5 Frame size for the word Tegadalay

Step 3: Windowing

This processing step is meant to window each individual frame so as to minimize the signal discontinuities at the beginnng and at the ending of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. Hamming window is the most commonly used window shape in speech recognition technology. Each frame has to be multiplied with a Hamming window in order to keep the continuity of the first and the last points in the frame [32].

If the signal in a frame is denoted by s (n), n = 0…N-1, then the signal after Hamming windowing is s (n)*w (n), where w (n) is the Hamming window defined by:

$$w(n, a) = (1 - a) - a\cos(2pn/(N-1)), \quad 0 \leqq n \leqq N-1$$

Or simply hamming window = frame * hamming (length (frame));

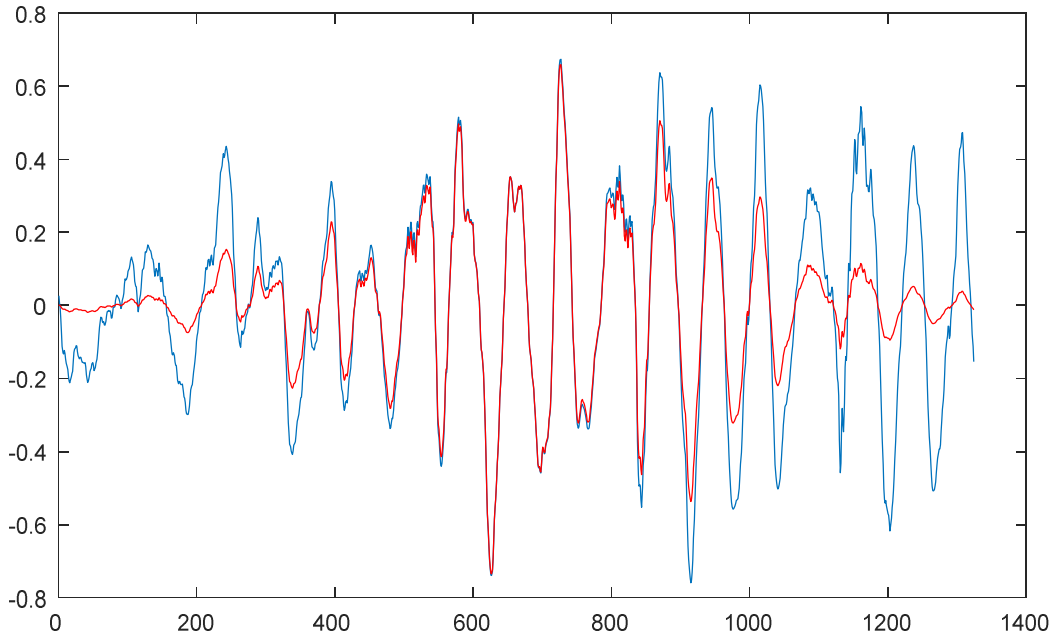


Figure 3-6 the Tegadalay Signal framing and Hamming Window red

Step 4: Fast Fourier Transform

A speech signal can be 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.

FFT= (abs (fft(frame)))

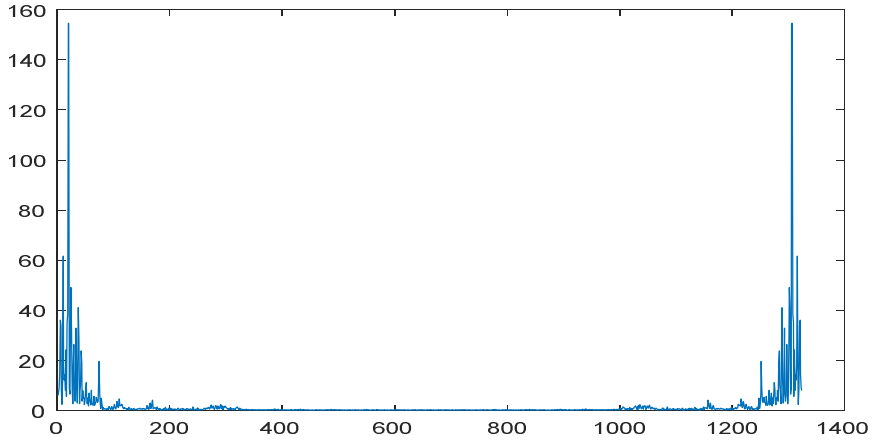


Figure 3-7 FFT of one frame signal

The result after this step is often referred to as spectrum or periodogram.

Step 5: Mel Filter Bank Processing

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. This method is an attempted to simulate the behavior of human ear system. The mapping frequency in hertz and the Mel scale is linear 1 kHz and logarithmic above 1 kHz. We multiply the magnitude frequency response by a set of 20 triangular bandpass filters to get the log energy of each triangular bandpass filter.

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

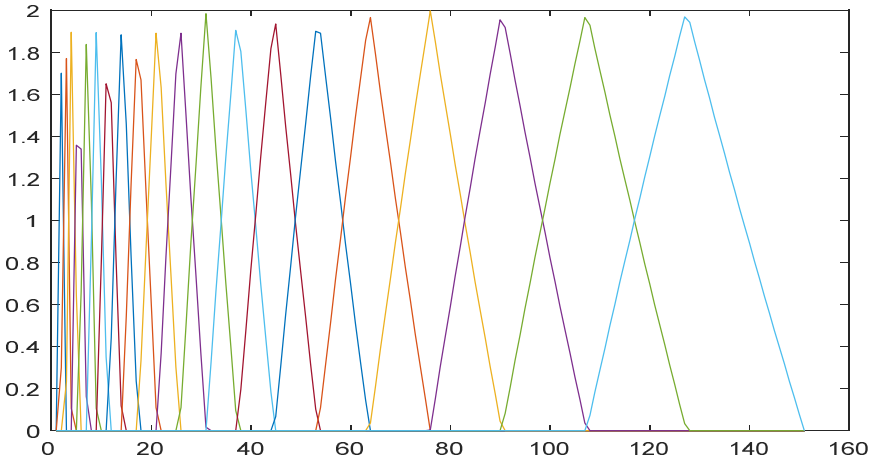


Figure 3-8 Mel filter bank For Tegadalay

Step 6: Discrete Cosine Transform

Finally, the inverse DFT (IDFT) of the log spectrum is calculated. The IDFT can be calculated as simple discrete cosines transform (DCT). This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of the coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of the acoustic vector [36].

The word Cepstrum is thus the spectrum of a spectrum.

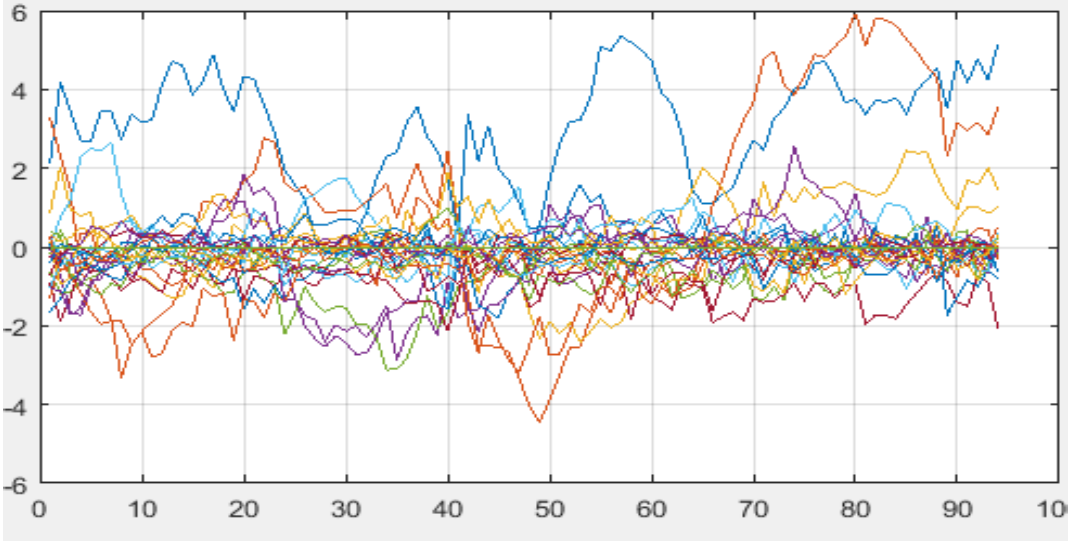


Figure 3-9 Cepstrum MFCC feature vector final stage

By applying the procedure described above, for each speech frame of around 30msec with overlap, a set of Mel-frequency cepstrum coefficients is computed. These are the result of a cosine transform of the logarithm of the short-term power spectrum expressed on a mel-frequency scale. This set of coefficients is called an acoustic vector which is derived from each frame. Therefore, each input utterance is transformed into a sequence of acoustic vectors. In the next section described feature matching pattern classification how those acoustic vectors can be used to represent and recognize the voice characteristic of the speaker.

3.8.5. Feature Matching Module of the speaker model

Once, after speaker template models are built and stored speaker database based on these features extraction; an utterance represented as a sequence of feature vectors speakers' vocal characteristics in the sound database. feature extraction made it's going to feed to the feature matching model of pattern comparison to involves identification of the unknown speaker by

comparing the extracted features from his/her voice input with the ones from a set of known speakers stored database template reference models. For the pattern comparison, neural networks (NN) were used. Many architectures or configurations are possible for the neural networks. The architecture selected for this system was the multi-layer perceptron pattern recognition specialized version of the feed-forward neural network. Levenberg-Marquardt back-propagation was used to train the neural network for learning algorithm [33].

The patternnet network is a specialized version of the multilayer feed-forward neural network architecture. The network was trained using the trainlm function, the training algorithm start via initializing weights, biased, then summing the inputs multiply weights then after input training matrix is applied to the network pass it to transfer function to extract the scaled output value. This output is then compared to the target matrix and assigns values to the different classes. As the training algorithm continues, the weights in the network are adjusted in order to minimize any errors in recognition using back-propagation. The training iterations continue, if not satisfied with the goal till to reaching best accuracy.

The following describe the specification of the design consist the following elements:

1. **The Input Layer**: The inputs for the neural network are matrices 26 feature vector and the number samples frame sizes for individual speaker are varied for a single utterance. The arrangement for the input vector was done so that the data is carefully divided by the training process done in MATLAB into 70% for training, 15% for validation and 15% for testing sets. The function selected to implement the mentioned division function was dividerand.

2. **The hidden layer**: 12 and 25 neurons with tan-sigmoid transfer functions were selected for the architecture. Tan-sigmoid found to provide better training results than other transfer functions.

3 **The output layer**: the number of output layer gives is depends on the number of classes stored in the database associated with class ID number and name. The transfer function for the output layer was selected to be tan- sigmoid again. After this it produce the desired output based on the supervise pattern recognition best closed on or most likelihood class ID were displayed with its optional identifier name as output.

4. **Building of the network**: The specialized feed-forward type of pattern recognition neural network is declared using the function: patternnet (12,'trainlm') or patternnet (25,'trainlm'). Inside

this function, the input vector, the output vector, the number of hidden neurons, the training algorithm and transfer function used in each layer are specified show on appendix C.

5. **Training Algorithm**: The training algorithm selected for this network was the Levenberg-Marquardt back propagation (trainlm). trainlm is a network training function that updates weight and bias states according to Levenberg-Marquardt optimization. trainlm is often the fastest backpropagation algorithm in the toolbox, and is highly recommended as a first choice supervised algorithm, although it does require more memory than other algorithms.

6. **Training Parameters**: Trainglm has many training parameters such as show on appendix C:

- Learning Rate: the Learning rate for the training was set to 0.01.
- Performance goal: the goal was set 0.0
- Minimum Gradient: 1.00e-07
- Echo initialed: 1000;
- Division method: 'dividerand'

In general, the designed the MLP neural network architecture had 26 input nodes, 25 hidden neurons, and 10 output neuron or 26 input nodes, 12 hidden neurons, and based on a selective number of speakers output neuron. The parameters for training were adjusted through observing the training process.

3.9.Programming Language Used and Justification

The programming language used to develop speaker recognition based on neural networks is matrix laboratory (in abbreviation MATLAB) version 9.0.0.341360 (R2016a). MATLAB is a fourth-generation high-level programming language and interactive environment for numerical computation, visualization, and programming. MATLAB is developed by MathWorks.

MATLAB is rich in built-in signal processing toolbox Signal Processing Toolbox provides functions and apps to generate, measure, transform, filter, and visualize signals. The signal processing toolbox to analyze and compare signals in time, frequencies, and time-frequency domains, identify patterns and trends, extract features, and develop and validate custom algorithms to gain insight into your data.

Neural Network Toolbox provides algorithms, functions, and apps to create, train, visualize and simulate neural networks. It helps to perform classification, regression, clustering, time-series forecasting, and dynamic system modeling and control.

The main reason researcher selected based on the two necessities and basic feature of MATLAB for developing are toolbox of signal processing and neural networks in addition to this in general it has a vast library, interactive environment, built-in graphics for visualizing data and tools for design, ability to access the PC microphone direct sound record and with custom graphical user interfaces.

MATLAB Neural Network Toolbox provides tools for designing, implementing, visualizing, and simulating neural networks. It supports feedforward networks, radial basis networks, dynamic networks, self-organizing maps, and other proven network paradigms.

The code for the project was written in MATLAB script m-files and for the speakers Sound Database it has saved and 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 Graphical User Interface (GUI).

3.10. Summary

In this chapter, we were discussed on the methodologies for speaker recognition by described the procedure step by steps. In the first case, the research designed for the speaker identification was quantitative study specifically follows experimental approach by exploratory data collection of the data sources corpus. Preparation and selecting sound recorded corpus are essential for the thesis collected in three ways direct recording by praat software, recorded from a phone conversation and directly recording live PC microphone. Sampling technique was used in the judgmental sample. The prototype of speaker recognition did in three phases those are preprocessing such as unvoiced sound detected, feature extraction signal processed by MFCC and finally pattern classifier based on artificial neural networks were applied.

CHAPTER FOUR

4. PRESENTATION, ANALYSIS, AND INTERPRETATION OF EXPERIMENTAL RESULT DATA

4.1. Introduction

In this chapter presented experimental results with their discussion and interpretation what they have been produced at each step of feature extraction and training of neural networks pattern recognition. The results of the gathered data are presented in figures and tabular including a description of what the data are all about and make an interpretation of the results of the computations. In this chapter also experiment results to test various modules performance evaluation to determine the accuracy of results. In the last chapter tried to show this study compare with previous related works. The experiments for different modules presented in different subsections below.

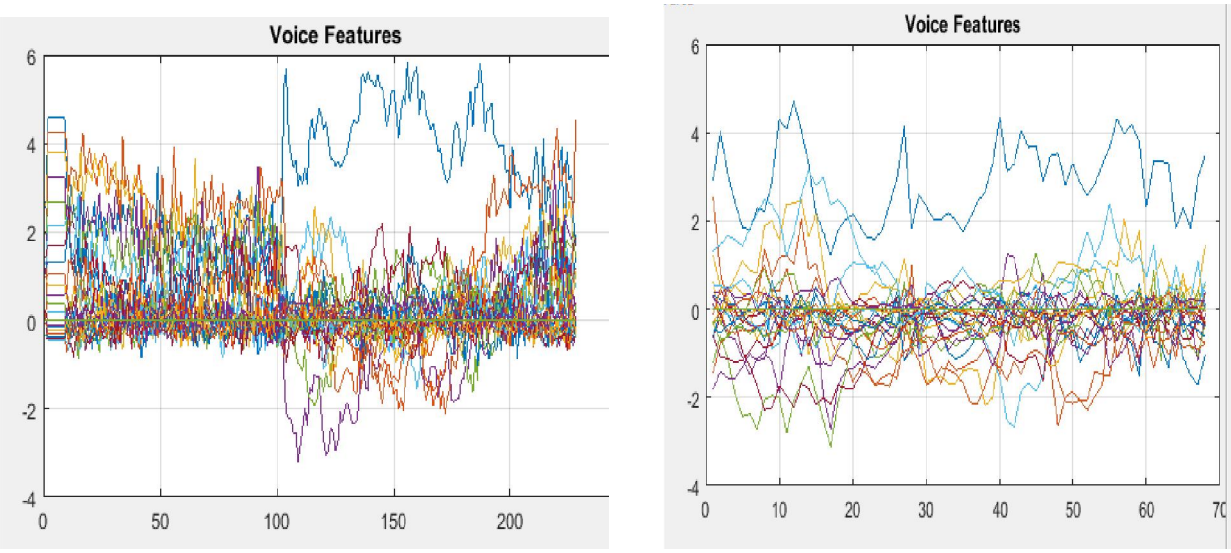
Table 4-1 Representation of speakers name into code

S.No	Speakers Name	Speakers name represented code
1	Ataklti	SAT
2	Bekuretsion	SBK
3	Birtukan	SBR
4	EqubayPhone	SEQP
5	Equbay	SEQ
6	Haile	SHI
7	Roza	SRZ
8	Selam	SSL
9	Tadelu	STD
10	Tirhas	STR

The speaker's name representation code derived from speaker name took the first alphabet from speaker S and from the first name took the first alphabet combine with additional one another alphabet for the sake of identifying codes' name.

4.2.Preprocessing

In figure 4-1 (a) and (b) from the speaker SAT shows the extracted vector of MFCC coefficient for the word tegadaly before EPD and after EPD. The end-point detection effect on the resultant MFCC coefficients appeared clearly, especially in the short-duration. The endpoint detection was followed by padding and uncleared pitch dark of resultant vector at the end and at the beginning of the utter speech as shown in the figure 4-1. Removing such silence part of voice has a great significance in the identification result process. This was done to follow a requirement for the neural network of having the same length as input vectors which are defined by 26 inputted vectors for all speakers. The coefficients were aligned to the center of the vector to be more suitable for the neural network.



(a) **(b)**

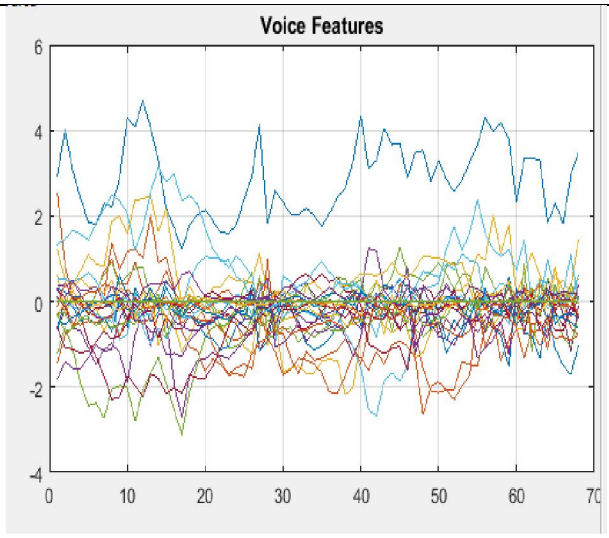
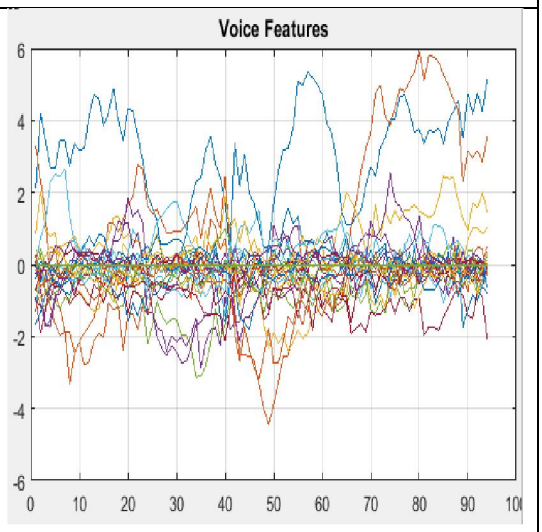
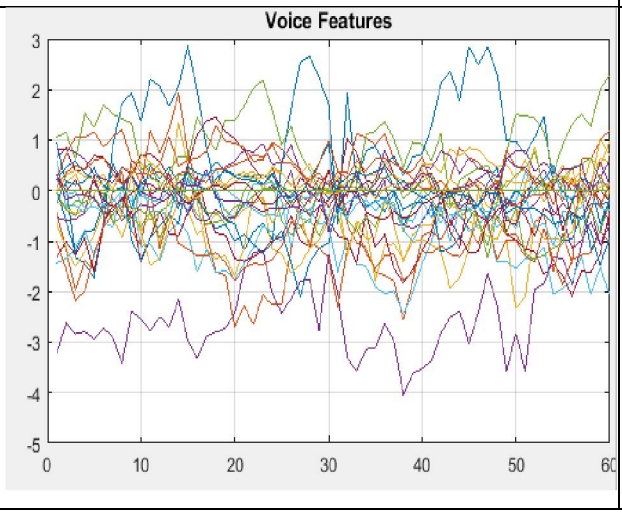
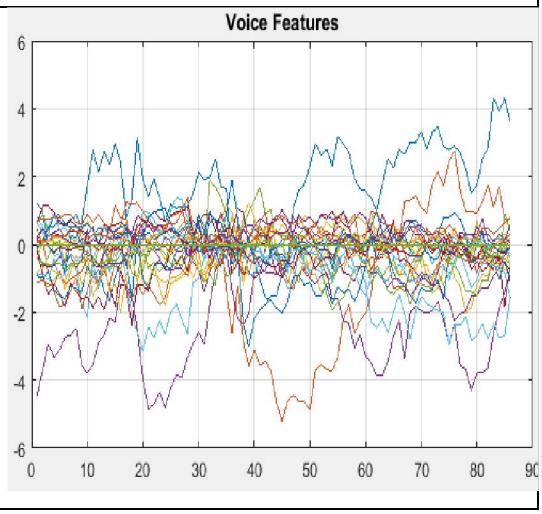
Figure 4-1 Voice features before (a) and after EPD (b)

4.3.Feature extraction

The following table 4-2 shows that, an experimental example of the final steps of the signal processing result of feature extraction. Those contain the voice features vector sometimes called coefficient of MFCC for two selected speaker. The table 4-2 contains plots of selected two speakers represent one from male (SAT) and one from female (STD). Each speaker represents the coefficient of the unique characteristics of voice feature (MFCC) of the word Tegadalay and Gazeta weyin final extracted stored in the sound database latter use for the matter of comparison

and feed to the neural network for testing to make a map comparison during feature classification.

Table 4-2 MFCC Coefficient of representation sample

	Voice feature vectors for tegadalay	Voice feature vectors for Gazeta weyin
Speaker SAT		
Speaker STD		

For instance, the number MFCC coefficients extracted and represented from speaker SAT are 26 extracted coefficients vector by 68 coefficients of frame samples for the word tegadalay and 26 by 94 for the phrase Gazeta weyin.

Assumed all speakers have their own MFCC coefficients (like in table 4-2) are indicates any coefficients extracted from the same word and phrase, from different speakers uttered the same which have different coefficients among the speakers. In addition to this, the general shape of

coefficient vector took from each frame was found to be different. These differences are the core idea caused by the identification. The feature extraction pattern classification model neural network's task is to map those sets of coefficients into the individual class of speakers.

	1	2	3	4	5	6	7	8	9	10
1	2.9412	2.5259	1.2040	0.3129	-1.2138	1.3244	-0.6371	-0.6528	-1.4300	0.6098
2	4.0134	0.8478	0.5158	-0.1077	-0.5940	1.4635	-0.1760	0.0590	-0.7594	-0.3515
3	3.0555	0.0789	0.3912	-0.6171	-1.1323	1.6651	0.1231	-0.1251	-0.6161	0.2655
4	2.4148	-0.0949	0.7266	-0.6904	-1.9152	1.5745	-0.3465	-0.4450	-0.1217	-0.0618
5	1.8577	-0.2456	1.0892	-1.0985	-2.4380	1.4415	-0.9627	-0.3876	-0.1576	-0.0800
6	1.7990	-0.3427	0.8802	-1.4856	-2.3680	1.8283	-1.2953	-0.5571	0.1754	-0.1509
7	2.2883	0.1355	0.8129	-1.1192	-2.7203	2.1599	-1.7422	-0.3934	0.7073	-0.6838
8	2.2174	1.3738	1.8862	-0.9818	-2.0531	2.4901	-2.2827	-0.7013	0.9730	-0.2765
9	2.8712	0.7347	2.0052	-1.1925	-1.9605	2.3696	-2.2265	0.0320	0.4314	0.3252
10	4.2948	1.1785	1.5800	-1.8847	-2.0123	2.0564	-1.7339	-0.5946	0.3998	0.2282
11	4.0817	1.2205	2.3716	-0.7775	-2.7991	1.2102	-1.9645	0.3186	0.9210	-0.3912
12	4.7021	1.0109	2.4051	-0.2274	-2.0652	1.7011	-2.2323	-0.0713	0.3119	-0.1071
13	4.0648	2.0319	2.4863	-1.0490	-1.6677	2.5657	-1.6920	-0.4692	0.0067	0.3126
14	3.2762	0.9082	1.6358	-0.2969	-1.3103	3.1476	-1.7899	-0.2592	0.2983	-1.0376
15	2.1775	1.0249	2.1658	-0.3167	-1.9115	2.8065	-2.1511	0.3344	0.7224	-0.4471
16	1.7424	0.3352	0.9817	-1.9309	-2.5294	2.9803	-2.0145	-0.0335	0.2481	-0.4903
17	1.2125	-0.6804	-0.0768	-2.7171	-3.1147	2.3620	-2.1470	-0.3780	-0.0643	-1.1787
18	1.7822	-1.0635	0.3732	-1.7253	-2.1055	2.4830	-1.7127	-0.1482	-0.1407	-0.8288
19	2.0277	-1.1559	-0.0032	-1.3934	-1.4662	2.2782	-1.7946	-0.4597	-0.0208	-0.3603
20	2.1390	-1.6419	0.1052	-1.3435	-0.8261	1.7414	-1.7935	-0.8097	-0.3983	-0.1012

Figure 4-2 Partial numeric representation of MFCC matrix in database speaker SAT

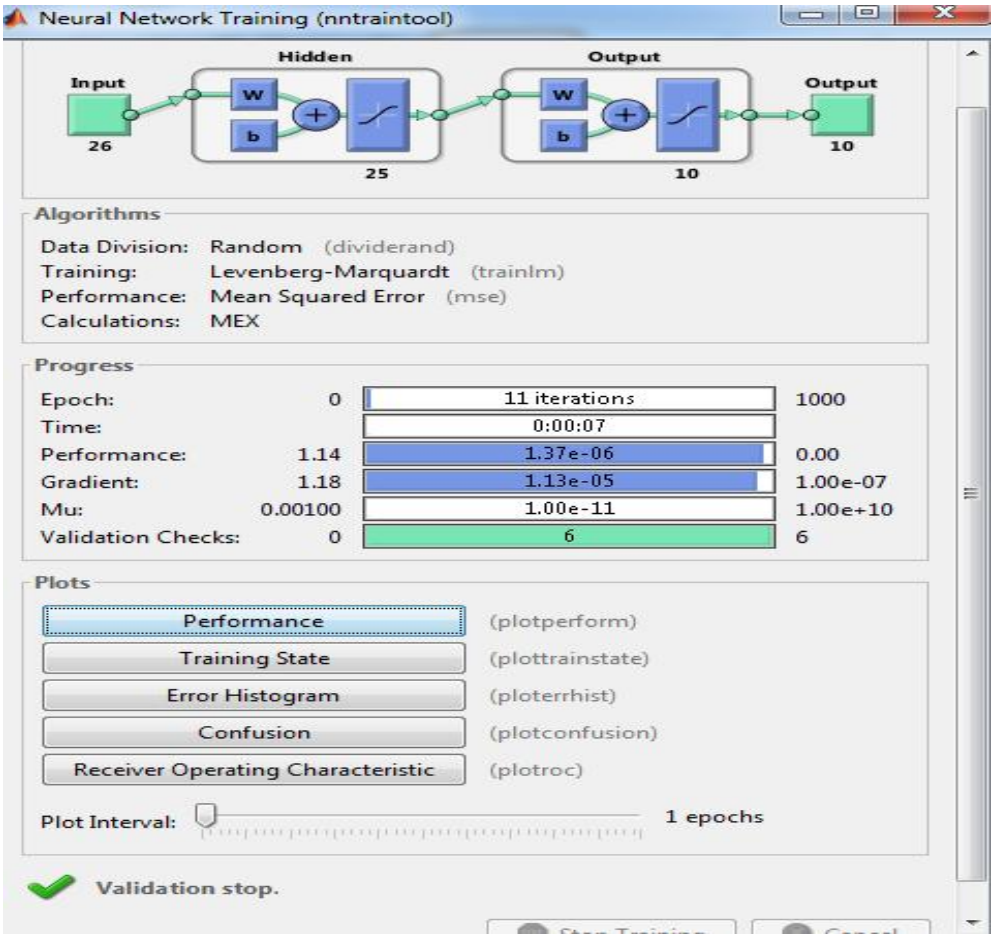
The above figure 4-2 MFCC coefficient matrix represented by each column represents one extracted coefficient vector for each row represents the coefficient of one frame samples of the speech stored in speakers' sound database.

4.4.Neural Network Model

After adjusting the training parameters, the set of inputs correspond with targets fed to the neural network in a batch mode. Training neural networks tool in MATLAB has a graphical user interface. This GUI provides some useful plots to observe the Neural Network Training see below figure 4-3. In the GUI showed the time of training, a number of (trial) epochs, performance, gradient and validation checkers. The training GUI also provides the plots of the performance, training state, ROC, and confusion.

The algorithms used for the nntraintool stated in the GUI. The trained network met the performance goal, hence the training stopped at epoch 11. The training took 7 seconds with a final gradient of 1.18. The GUI illustrated the architecture of the trained neural network; one

hidden layer with a tan-sigmoid activation function and an output layer with a tan-sigmoid activation function. The training algorithm used was levenberg-marquardtb Back-propagation. The performance measure used was Mean-square error and confusion matrix.



.Figure 4-3 Neural network Training tool Speaker recognition

The goal supervised pattern recognition is to classify the patterns into class categories. The classes here refer to individual speakers. Since the classification procedure in our case is applied to extracted features, it can be also referred to as feature matching. During the training session, researcher label each input speech with the ID of the speaker (1 to 10).

The following figure 4-4 is MLP designed with three layers of the feed-forward neural network training Graphical User Interface (GUI) as shown in figure 4-4. In the GUI, MLP pattern recognition neural network it has input layer 26 vector coefficient for each speaker, in the hidden layer selected 25 neurons and at the output layer classified based on the number of classes here is 10 classes.

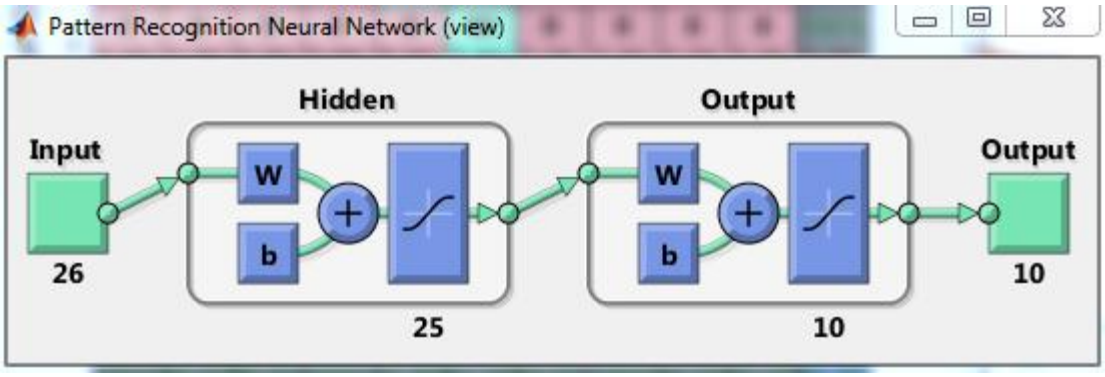


Figure 4-4 Pattern Recognition NN simulation result view

Figure 4-5 is the plot of the performance for the training, validation, and testing versus the epochs. The blue curve was for the training performance, the green curve was for the validation performance and the red curve was for the test performance. 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 0.06504 at epoch 5.

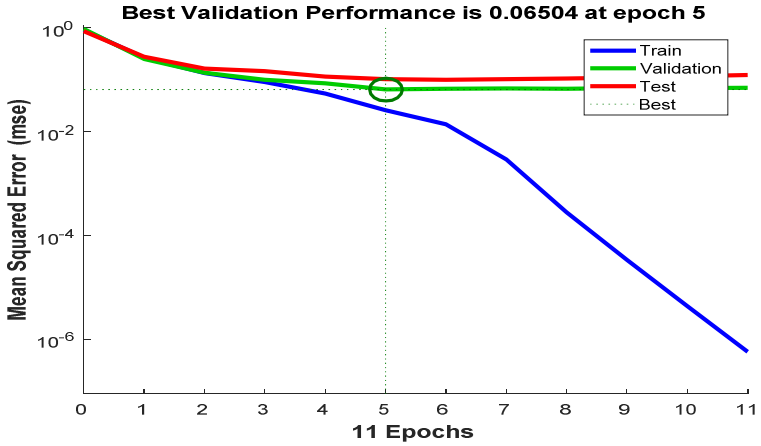


Figure 4-5 Performance plot

4.5.Performance Evaluation parameters

Evaluating machine learning algorithm is an essential part of any kind research such as speaker recognition using ANN. To evaluate the prototype is the following evaluation parameters were used.

4.5.1. Accuracy

Speaker recognition is one of the tasks for pattern recognition classification. The accuracy performance analyzed by arranging correctly and incorrectly identified samples. Neural networks

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

A true positive [TP] is the correct identification of a speaker, a false positive [FP] is the incorrect identification of a speaker, a true negative [TN] is the correct identification of the incorrect speaker, a 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.

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

Classification accuracy is the ratio of a number of correct predictions to the total number of input samples. From the confusion matrix result, the main diagonal is classification accuracy for each class was computed in percentage.

Accuracy = TP/Total number of classification

Most of the time's classification accuracy use for measure the performance of the model, however, it is not enough to truly judge neural network model. Now here cover different types of evaluation metrics available. Confusion matrix as the name suggests gives us a matrix as output and describes the complete performance of the model. A based on the multi-class classification with the built-in function plotconfusion (t, y) see appendix C. Plots confusion matrix which is a very concise graphical display of how your network misclassified things and shows percentages of correct/incorrect in each class as well as the total. T is the targets matrix and y is the computed output, which you can extract using y=net (P); for P input matrix.

Confusion matrix displays the number of a correct and incorrect prediction made by the model compared with the target classification in figure 4.6 and figures 4.7 the test data 10*10 where 10 is the number of classes for the speakers. Below are criterion functions to assess classifier performance.

4.5.2. Recall, Sensitivity, and TPR

Sensitivity commonly known as recall is another 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 gives 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).

Calculated as: recall= sensitivity=TPR= TP/ (TP+FN)

Where, TP and FN are the numbers of true positive and false negative predictions for the considered class. The false negative rate or (miss rate) is the proportion of positives, which yield negative test outcomes with the test.

False negative rate = FN/FN+TP = FN/actual positive or FNR=1-sensitivity (TPR)

4.5.3. Specificity and TNR

Similar to Sensitivity is the evaluation parameter of Specificity. It is a proportion of the number of true negatives overall actual negatives.

Calculated as specificity=TNR= TN/ (TN+FP)

Where, TN and FP are the numbers of true negative and false positive predictions for the considered class. The false positive rate (or "false alarm rate") is calculated as the ratio between the number of negative events wrongly categorized as positive (false positives) and the total number of actual negative.

False positive rate, false alarm = FP/ (TN+FP) or FPR=FP/actual negative = 1 – specificity

4.5.4. Precision

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

Where, the TP and FP are the numbers of true positive and false positive prediction of the considered class.

4.5.5. F1-score

The F_1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision and the recall. F1-score is the harmonic mean of precision and sensitivity

Its formula is 2*(precision * recall) / (Precision + recall)

F_1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

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.

Figure 4-6 shows us the rows correspond to the predicted class (Output Class) and the columns correspond to the true class (Target Class).

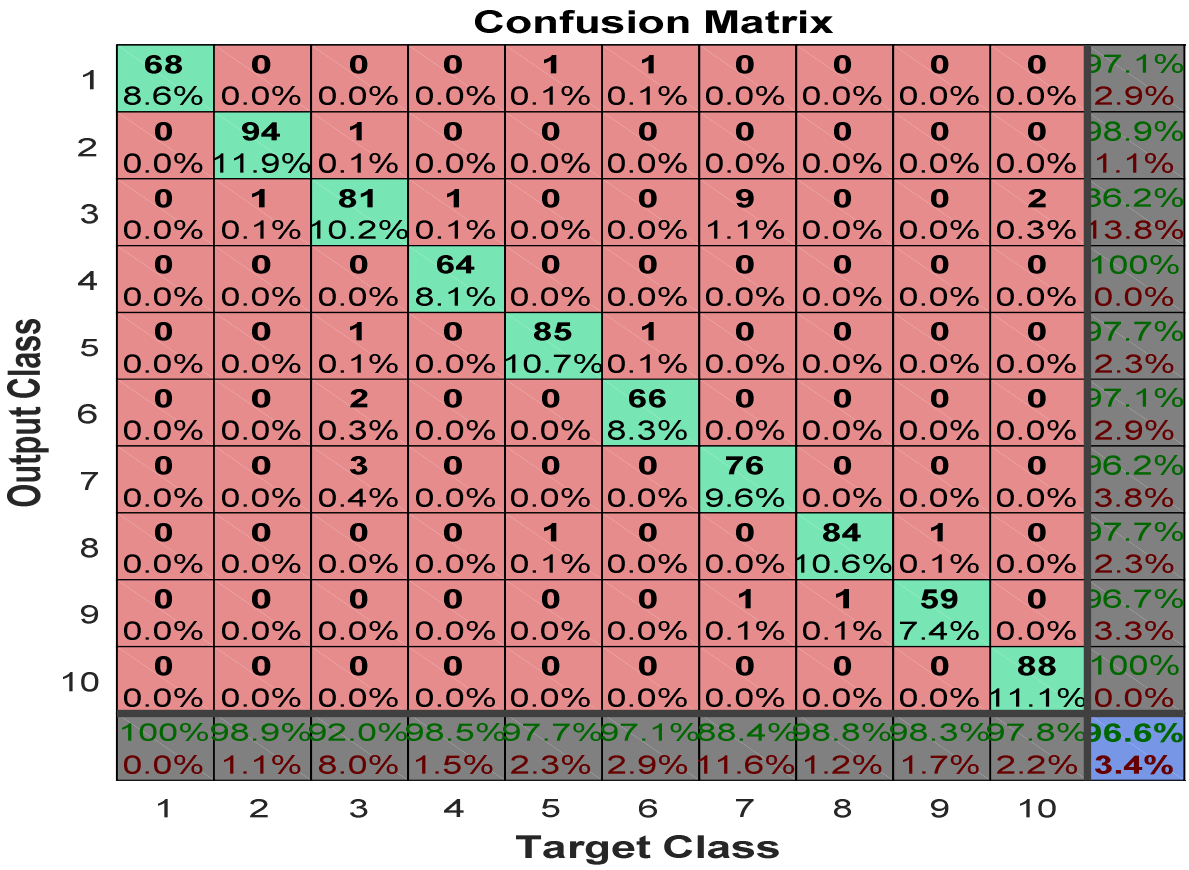


Figure 4-6 Confusion matrix for the word Tegadalay

From the figure, 4-6 shows us confusion matrix for the word Tegadalay displays the diagonal cells show the number of voice samples that were correctly classified, and the off-diagonal cells show the misclassified samples of voice. The blue cell in the bottom right shows the total percent

of correctly classified voice samples in green and the total percent of misclassified voice samples in red. The results show very good recognition.

The column on the far right of the above figure shows the percentages of all the examples 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 (in red color) rate respectively. The row at the bottom of the above figure shows the percentages of all the examples belongs to each class that is correctly and incorrectly classified respectively. 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.

From class 1 its accuracy is 100% highest accuracy from all classes which indicated all the voice samples are correctly classified and no predicted voice samples or misclassified samples in target class 1. From class 3 and class 7 the accuracy is 92.0 and 88.4 respectively. The identification rate of the classes 3 and 7 are less than others because of the recorded their voice utterance are at environmental noise influence on the classification and accuracy of the model. The variance range from highest accurately classified and the lowest classified is 11.6%.

Accuracy for the matrix is calculated by taking an average of the values lying across the main diagonal for individual classes and the final accuracy is at the right bottom.

The accuracy of model =96.6%

Classification predicted accuracy = 99.3%

Accuracy is the percent of correct classifications and Error rate is the percent of incorrect classifications.

Error Rate = 1 – accuracy.

Error rate =3.4%

The accuracy has been done from the testing set and the confusion matrix divided into four categories' those with their accuracy percentage are [see table 4-7]: Training class =98.9%, Testing =89.9%, Validation =92.4% and Overall accuracy is 96.6%.

Table 4-3 Metrics value from confusion matrix for each class for the word Tegadalay

Metrics / Speakers Name	Total samples of classes	TP	FN	FP	TN
Class 1 SAT	68	68	0	2	722
Class 2 SBK	95	94	1	1	696
Class 3 SBR	88	81	7	13	691
Class 4 SEQP	65	64	1	0	727
Class 5 SEQ	87	85	2	2	703
Class 6 SHI	68	66	2	2	722
Class 7 SRZ	86	76	10	3	703
Class 8 SSL	85	84	1	2	705
Class 9 STD	60	59	2	2	729
Class 10 STR	90	88	2	0	702
Total and average	**792**	**96.6%**			

From the above table 4-3, have four possible results which indicate how to classify the samples sound frames versus with four-term parameters. For instance, class 1 SAT, has total 68 sound sample frames stored in sound database extracted from his voice. From those, the truly classified samples are the actual class 1 predicted correctly 68 (TP). The 2 (FP) frame samples are classified incorrectly or predicted as class 1. The 0 (FN) frame samples of actual class 1 are predicted incorrectly by other classes. The 722 (TN) frame sampled neither actual class nor predicted class which means truly predicted incorrectly from the total samples 792.

Based on this principle below table 4-4 each class is calculated by their confusion matrices and different metrics parameters of accuracy for the word Tegadalay.

Table 4-4 Computed the accuracy of model performance for the word Tegadalay

Speakers name / Metrics	Class 1 SAT	Class 2 SBK	Class 3 SBR	Class 4 SEQP	Class 5 SEQ	Class 6 SHI	Class 7 SRZ	Class 8 SSL	Class 9 STD	Class 10 STR
Accuracy of the model	8.6%	11.9%	10.2%	8.1%	10.7%	8.3%	9.6%	10.6%	7.4%	11.1%
Classification accuracy	99.7%	99.7%	97.5%	99.9%	99.5%	99.5%	98.4%	99.6%	99.5%	99.7%

Precision	97.1%	98.9%	86.2%	100%	97.7%	97.1%	96.2%	97.7%	96.7%	100%
Recall	100%	98.9%	92.0%	98.5%	97.7%	97.1%	88.4%	98.8%	98.3%	97.8%
F1-score	98.5%	98.9%	89.0%	99.2%	97.7%	97.1%	92.6%	98.2%	97.5%	98.9%
FPR	0.3%	0.1%	1.8%	0%	0.3%	0.3%	0.4%	0.3%	0.3%	0%
TPR	100%	98.9%	92.0%	98.5%	97.7%	97.1%	88.4%	98.8%	98.3%	97.8%
FNR	0%	1.1%	8.0%	1.5%	2.3%	2.9%	11.6%	1.2%	1.7%	2.2%
TNR	99.7%	99.9%	98.2%	100%	99.7%	99.7%	99.6%	99.7%	99.7%	100%
Sensitivity	100%	98.9%	92.0%	98.5%	97.7%	97.1%	88.4%	98.8%	98.3%	97.8%
Specificity	99.7%	99.9%	98.2%	100%	99.7%	99.7%	99.6%	99.7%	99.7%	100%

Figure 4-7 below shown the phrase Gazeta weyin follow the same principle as the above confusion matrixes for the word Tegadalay.

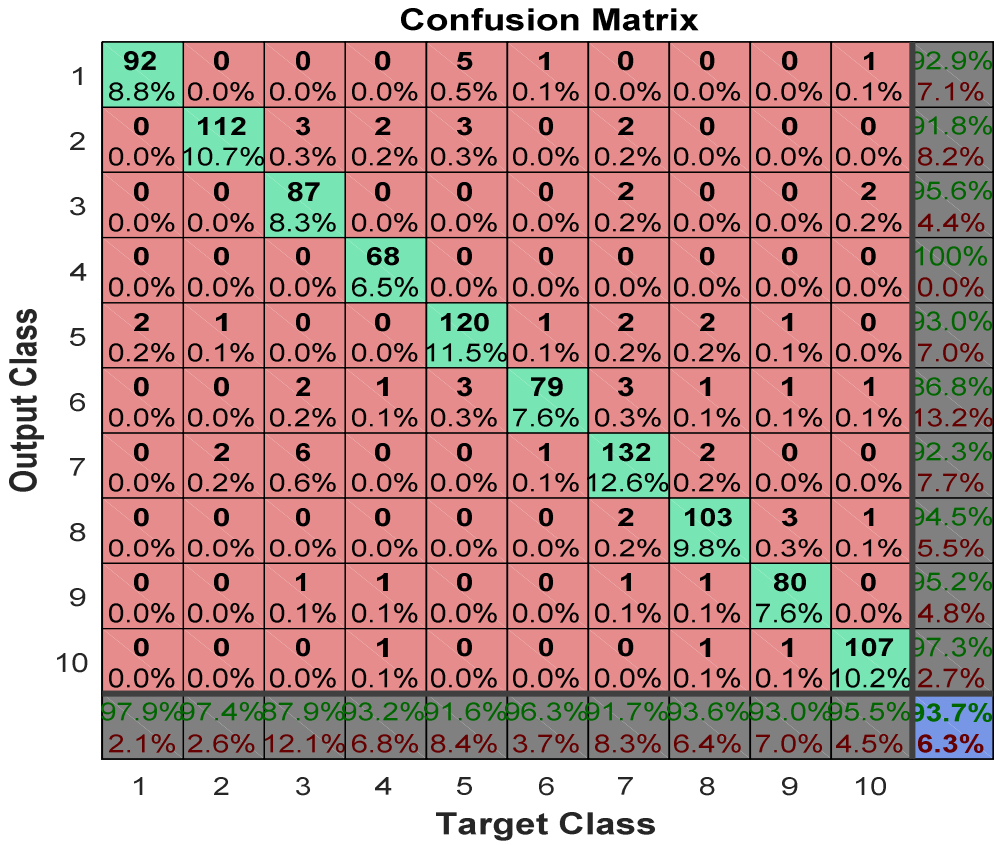


Figure 4-7 Confusion matrix for the phrase Gazeta weyin

From the figure, 4-7 shows us confusion matrix for the phrase Gazeta weyin the same procedure followed with the word tegadalay above figure 4-6 the diagonal cells show the number of voice samples that were correctly classified, and the off-diagonal cells show the misclassified samples of voice. The column on the far right of the above figure shows the precision (or positive predictive value in green color) and false discovery (in red color) rate respectively. The row at the bottom of the above figure shows the recall (or true positive rate in green color) and false negative rate (in red color), respectively. The cell in the bottom right of the plot shows the overall accuracy.

From class 1 its accuracy is 97.9% highest accuracy from all classes which indicated 92 out of total sample 94 the voice samples are correctly classified and 2 voice samples misclassified samples as predicted class 4. From class 3 and class 7 the accuracy is 87.9% and 91.7% respectively. The identification rate of the classes 3 and 7 are again less than others because of the recorded their voice utterance are at environmental noise influenced even in the phrase utterance on the classification and accuracy of the model. The variance range from highest accurately classified and the lowest classified is 10%.

Accuracy for the matrix can be calculated for each class at main diagonal for individual classes and the final accuracy result is at the right bottom**.**

The accuracy of model =93.7%

Classification predicted accuracy = 98.74%

Accuracy is the percent of correct classifications and Error rate is the percent of incorrect classifications.

Error Rate = 1 – accuracy.

Error rate =6.3%

The accuracy has been done from the testing set and the confusion matrix divided into four categories' those with their accuracy percentage are [see table 4-7]: Training class =99.9%, Testing =77.7%, Validation =80.9% and combined Overall accuracy is 93.7%.

Table 4-5 Metrics value from confusion matrix for each class for the phrase Gazeta Weyin

metrics / **Name**	**A total sample of classes**	**TP**	**FN**	**FP**	**TN**
Class 1 SAT	94	92	2	7	945
Class 2 SBK	115	112	3	10	921

Class 3 SBR	99	87	12	4	943
Class 4 SEQP	73	68	5	0	973
Class 5 SEQ	131	120	11	9	906
Class 6 SHI	82	79	3	12	952
Class 7 SRZ	144	132	12	11	891
Class 8 SSL	110	103	7	6	930
Class 9 STD	86	80	6	4	956
Class 10 STR	112	107	5	3	931
Total and average	**1046**	**93.7**			

As we can from the above table 4-5 consists the matrixes value of four metrics parameter results with their classification of the phase samples. This table 4-5 follows the same calculation principle with table 4-3 above.

Table 4-6 Computed the accuracy of model performance for the phrase Gazeta Weyin

Name / Matrix	Class 1 SAT	Class 2 SBK	Class 3 SBR	Class 4 SEQP	Class 5 SEQ	Class 6 SHI	Class 7 SRZ	Class 8 SSL	Class 9 STD	Class 10 STR
Classification accuracy	8.8%	10.7%	8.3%	6.5%	11.5%	7.6%	12.6%	9.8%	7.6%	10.2%
Accuracy of the model	99.1%	98.8%	98.5%	99.5%	98.1%	98.6%	97.8%	98.8%	99.0%	99.2%
Precision	92.9%	91.8%	95.6%	100%	93.0%	86.8%	92.3%	94.5%	95.2%	97.3%
Recall	97.9%	97.4%	87.9%	93.2%	91.6%	96.3%	91.7%	93.6%	93.0%	95.5%
F1-score	95.3%	94.8%	91.6%	96.5%	92.3%	91.3%	92.0%	94.0%	94.1%	96.4%
FPR	0.7%	1.1%	0.4%	0%	1%	1.2%	1.2%	0.6%	0.4%	0.3%
TPR	97.9%	97.4%	87.9%	93.2%	91.6%	96.3%	91.7%	93.6%	93.0%	95.5%
FNR	2.1%	2.6%	12.1%	6.8%	8.4%	3.7%	8.3%	6.4%	7.0%	4.5%
TNR	99.3%	98.9%	99.6%	100%	99.0%	98.8%	98.8%	99.4%	99.6%	99.7%
Sensitivity	97.9%	97.4%	87.9%	93.2%	91.6%	96.3%	91.7%	93.6%	93.0%	95.5%
Specificity	99.3%	98.9%	99.6%	100%	99.0%	98.8%	98.8%	99.4%	99.6%	99.7%

4.6.Overall Result Summary

Table 4-7 Result Summary

Criteria of result And their parameters		Word Tegadalay			Phrase Gazeta Weyin		
		When 12 neurons used in hidden layers	When 25 neurons used in hidden layers		When 12 neurons used in hidden layers	When 25 neurons used in hidden layers	
5 speakers		99.5%	99.7%		97.0%	98.2%	
10 Speakers		95.5%	Training confusion Matrix	98.9%	88.6%	Training confusion Matrix	99.9%
			Validation Confusion Matrix	92.4%		Validation Confusion Matrix	80.9%
			Testing Confusion Matrix	89.9%		Testing Confusion Matrix	77.7%
			Overall Confusion matrix	**96.6**%		**Overall Confusion matrix**	**93.7**%
Based on Gender	5 Males	96.2%	97.9%		96.3%	98.2%	
	5 Females	95.0%	95.9%		95.3%	96.2%	

From the table 4-7 result tested by dividing the number of the speaker in number, gender and increase the number of hidden neurons, reset the initialed neural network parameters, increasing and decreasing the input training vector to new values and trained again has significantly changed the result and performance of the model. In order to determine an optimal number of hidden neurons followed the approaches. At first half of input voice feature vectors and equal to the input voice feature vectors but the number of output depends on the number of speakers.

From the above table observed the number of the speaker are 5 in the hidden layer used 12 neurons increase to 25 increase the accuracy of the model 0.2% for the word and 1.2% for the

phrase respectively performed. The numbers of a neuron are increasing importance for the variability of vocabulary has a great role to have a better accuracy result. From here in all aspects from the table 4-7 above the number of neuron increase, it increases the accuracy of the model effect on the performance of the model. Actually, it increases the response times during train the networks when increasing the number of the neuron. It has been observed that increasing the number of neurons in general increases the efficiency.

The other factor parameter is the number of speakers. When the numbers of speaker rise from 5 to 10 speakers it reduces the accuracy of the model by 3.1% for the word and 4.5% for the phrase respectively. It's known in speaker recognition the number of speaker increase always the number of accuracy decrease. This is because, as the number of templates increases, the classifier's accuracy in matching the feature array with the appropriate template declines. In addition to this, as the number of speakers increases, the probability of having similar templates increases. This increment in similarity makes the system to pass a wrong decision on the identity of the individuals.

The size and variability of vocabulary are also affected and decreased the accuracy when single word compared with the phrase by 2.9% from the experimental result. In the last, the other thing which is affected and observed from the result is environmental noise and sound energy. Although females have a higher pitch than males during recorded the level of energy are affected less silence and in the less noisy environment caused consequence of lower accuracy and identification rate than males.

Phone-based speeches may behave differently not very consistent output. The channel and the telephone equipment may affect speech quality. Testing performs the final model can be from the testing set and training set but researcher preferred from the testing data portion. Performance on the training set is a good but not good indicator to generalize the system performance.

4.7.Comparison with other work

Sometimes comparison works with previous research is needed especially did not perform speaker recognition studies for the local language Tigrigna. it is important to compare this work with the one performed for another language. One from Ethiopian national language has been worked on Amharic based on another model a little bit similar need to compare with it and even worked abroad indeed. The reason why researcher chooses this research work is that the speech

features, as well as the purpose of modeling techniques used, are similar. Below table 4-8 is summary of the comparison this study and with previous research work.

Table 4-8 Comparison previous work and this study

Parameters	**A previous study [1]**	**A previous study[9]**	**Previous Study[7]**	**This study**
No. of Speakers	50 Speakers, 500 utterances	9 Speakers, 180 utterances	15 speakers, undefined	10 Speakers, 40 utterances
Implementation tool	MATLAB	MATLAB	MATLAB	MATLAB
Features extracted and used	MFCC	LPC	MFCC	MFCC
Speech Language	Amharic	Thai	Hindi	Tigrigna
Classification algorithm/s used	VQ and GMM	ANN	VQ and GMM	ANN
Approach Used	Text Independent	Text Dependent	Text Dependent	Text Dependent
Evaluation metric/s used	Accuracy and subjective evaluation	Subjective evaluation	Accuracy based on subjective	Accuracy and Confusion matrix
Best result Obtained	Using VQ 74.2 % and using GMM 84.3 %	95%	Using VQ 85.49 % and using GMM 94.12 %	96.6% for a word and 93.7 for phrase

Speaker recognition for different languages is still challenging for researchers. the research seeks more accuracy to add to the performance of the speaker identification system in order to achieve the more specified recognition of the speaker. Achieving the goal requires a focus on the methodologies followed for both feature extraction and feature match with different parameters. Besides, evaluation of the system should be more reliable to accomplish precise evaluation. From the comparison, table 4.8 shown previous studies [1] the identification rate was increased

by 12.3%, a previous study [9] the identification rate is increased by 1.6 % and a previous study [7] the identification rate is increased by 2.5%. However, previous research works were reported on different corpora, some of them were private database while others were collected by recording the sample voices. Researcher compares its performance without considering the exact subset division, the quality microphone of sound recorder was used with their specific parameters and they followed different evolution methods.

Generally, the result clearly shows that still better work to achieved performance based on accuracy and excellent identification rate. Actually, only accuracy rate could not be considered for model performance such as response time, security and other issues should be included.

CHAPTER 5

5. CONCLUSIONS AND RECOMMENDATIONS

5.1.Introduction

This chapter has presented appropriate conclusions from the experimental results, a conclusion based on the investigations, contributions and forward recommendations to those who want study in the area of speaker recognition any other necessary for the study.

5.2.Summary of Findings and conclusions

In this study, the problem of speaker identification has been extended from unknown person's speech utterances to identifying a person's based on the voice characteristics. Based on the experimentally obtained results closed set text-dependent speaker recognition system research studied the possibility of using a combination of MFCC at the stage of feature extraction and ANN at the stage of pattern comparison and classification for speaker recognition systems has been developed. The testing and training of the prototype occurred in two ways, real-time or with recorded test samples. Test speaker model from testing dataset gives a better estimate of the classification error probability. Whereas testing the model from training dataset gives a better generalization. But never recommend evaluating performance from training datasets. It makes the conclusion would be optimistically biased.

In this research task consideration, many factors including sex, age, changes in dialect, an ideal setting less background noise, before and after EPD, with background noise, from a telephone conversation, and real-time by microphone effects have been considered. Based on these factor parameters were affected by the classification of the sound frame samples in the as shown confusion matrixes tables 4-4 and 4-6. This factor proved reduction on the result of identification on sensitivity since when increases the FN values it decreases the correctly and perfectly predicted classification. In addition, these factors forced to the sample of voice misclassification it caused to happen to increase the FP in order to less precise classification.

The other hand which is affected the model performance result is silence at the beginning and end of the speech. From the figure 4.1 (a) and (b) it's clearly distinguishing the effects from the resulting figure by making inconsistency results of inputs feature vector.

Generally, noise affects the system and decreases its performance. The erroneous results are due to the fact that the recording conditions are not optimal (background noise, echo etc), quality of microphones, the emotional state of the speaker, number of speaker and size of vocabularies, silent parts of speech, overall signal energy etc. can affect the recognition process. But performance can be optimized by deeply considering works on those factors. We got higher reorganization rate with the system average identification accuracy 96.6% for the word tegadalay and for the phrase Gazeta weyin is 93.7% for 10 speakers. Although, decrease the performance of the model were individually tested the word Tegadalay and Gazeta weyin are perfectly all speaker are correctly identified and classified as their class from the audio files and real time. The evaluation was made based primarily on the three standard evaluation parameters in classification tasks: confusion matrices based accuracy, sensitivity, and specificity.

This approach gives a good result for text-dependent speaker recognition. In some case, it gives 0% false rejection error especially when fewer speakers were stored in the speaker sound database. But if the number of the speaker is increased, it may cause an increase in the false identification error. However, this system has to be applied for a small number of speakers' grouped in 5 and 10 speakers rather than in large numbers.

Point to be considered and underlined, the researcher needs to emphasizing on neural network trained the whole result percentage of accuracy of the model could be varied may not be a consistent result. The main causes were weight, bias values and different random divisions of data into training, validation, and test sets. As a result, different neural networks trained on the same problem can give different outputs for the same input. To ensure that a neural network of good accuracy has been found, after retrain several times that's why researcher provides selected the best and optimal accuracy.

The architecture selected for the neural network in this thesis, which is a pattern recognition feed-forward neural network with one hidden layer containing 25 neurons and an output layer containing 10 neurons, inputted 26 coefficients vector was found to be effective and suitable for the identification process.

Finally, we concluded from the positive results collected and obtained: speaker recognition based on the neural networks speaker model showed good promising results for text-dependent speaker identification, although, the system has made error rate 3.4% for the word and 6.3 % for the phrase, its performance, perfectly identification, and efficiency have outshined.

5.3.Contributions

- The researcher designed and implemented ANN model based voice identification system on the limited studies area and has not been done before for Ethiopian language such as Tigrigna language represent a novel contribution to the state of the art into speaker recognition field.
- The researcher tried to design to connect the real-time and the tests set to train and test the model simultaneously.
- The researcher tried to show the variance of performance factors where sound recorded from the noisy, silent environment and size of vocabularies.
- Finally, research contributions improving the performance from the previously studied although, they were used different corpora and followed different evolution methods.

5.4.Recommendations and Future Works

After achieving the stated objectives of the thesis, future plans to further exploit its capabilities and make full use of its potential have been laid out below. The researcher would like to recommend the following for future works.

- Spoofing is a serious concern to enhance security for speaker recognition platform designers, anti-spoofing will be a concern to move up in the commotion
- Use deep neural network for speaker identification by using large vocabulary with multiple layers are trained with large amounts of the node, neurons produce state-of-the-art results on many pattern recognition tasks and capable of capturing long-distance relations.
- Speaker Identification based on Gender, age and dialects base on pronunciation.
- horizontal improvements for speaker recognition, in the form of expanding currently available aspects of the project such as using larger sets of speakers, speakers spoofing, vertical improvements, better end-point detection algorithm and using the wavelet transformability to de-noise the signals in order to improve performance.
- The voice of speakers may change due to many different reasons such as the health of the speaker, the emotional status of the speaker, multilingual speakers, and to attempt to guard against voice mimicry. One way to handle this problem is to have models, which constantly adapt to changes.
- Future research should also focus on the comparison of this study like HMM, SVM, and others by increasing the numbers of the training and testing datasets.

References

[1]. Aykefam Azene "Text-Independent Speaker Identification for the Amharic Language" Master's Thesis, Bahir Dar Institute of Technology, Bahir Dar University, Jan, 2015

[2]. Xiaoguo Xue "Joint Speech and Speaker Recognition Using Neural Networks" Bachelor's Thesis, Electrical Engineering, Vaasa, May 9, 2013

[3]. R.V Pawar, P.P.Kajave, and S.N.Mali "Speaker Identification using Neural Networks" *World Academy of Science, Engineering and Technology*, vol. 2, page 31-35, 2005

[4]. Bich Ngoc Do "Neural Networks for Automatic Speaker, Language and Sex Identification" Master Thesis, Faculty of Mathematics and Physics, University of Groningen, Nov, 2015

[5]. Alexander Murphy "Implementing Speech Recognition with Artificial Neural Networks", Master Thesis, Department of Computer Science, Algoma University, Apr 11, 2014

[6]. Hafte Abera, "Hidden Markov Model Based Large Vocabulary, Speaker Independent, and Continuous Tigrigna Speech Recognition". Master's Thesis, Department of Information Science, Addis Ababa University, Nov, 2009.

[7]. Ankur Maurya, Divya Kumar and R.K.Agarwal , "Speaker Recognition for Hindi Speech Signal using MFCC-GMM Approach", *6th International Conference on Smart Computing and Communication*, Kurukshetra, India, ICSCC 2017, 7-8, Dec, 2017

[8]. Rosistem build your business. [Online]. Available: http://www.barcode.ro/tutorials/biometrics/voice.html, Date accessed Apr 15, 2018.

[9]. Chularat Tanprasert, "Text-dependent Speaker Identification Using Neural Network on Distinctive Thai Tone Marks", *Technical journal*, vol. 1, no. 6, Feb, 2000

[10]. J. R. Deller, J. H. L. Hansen and J. G. Proakis, "Discrete-Time Processing of Speech Signals, Piscataway (N.J.)", *IEEE Press*, 2000.

[11]. Pradeep. CH, "Text Dependent Speaker Recognition Using MFCC and LBQVQ", National Institute of Technology, Rourkela, 2007

[12]. David Sierra Rodríguez, "Text-Independent Speaker Identification", AGH University Of Science and Technology Krakow, Kraków, 2008

[13]. Homayoon Beigi, "*Fundamentals of speaker recognition"*, Yorktown Heights, NY, Springer Science+Business Media, LLC 2011

[14]. Girmay Berhane "The phonology of Tigrigna". Master's Thesis, Addis Ababa University, Addis Ababa, 1983

[15]. Tigrigna language. [Online]. Available:
http://en.wikipedia.org/wiki/Tigrinya_language, date accessed on Mar 5, 2018

[16]. H. Gish and M. Schmidt, "Text Independent Speaker Identification", *IEEE signals processing Magazine*, Vol. 11, No. 4, pp. 18-32, 1994.

[17]. Pablo Zegers, "Speech Recognition Using Neural Networks", the University of Arizona, 1998

[18]. SBS Radio [Online]. Available:
http://www.sbs.com.au/yourlanguage/tigrinya/, date accessed on Mar 5, 2018

[19]. Rodman, R. D. "Computer Speech Technology". Norwood: Artech House, Inc. 1999

[20]. Xiaoguo Xue, "Joint Speech and Speaker Recognition Using Neural Networks", Novia University of applied science, May 9, 2013

[21]. Dipen Nath and Sanjib Kr. Kalita, "Feature Selection Method for Speaker Recognition using Neural Network", *International Journal of Computer Applications (0975 – 8887),* Volume 101– No.3, September 2014

[22]. Dan Bakmand-Mikalski, "Speaker identification", Master thesis, Technical University of Denmark (DTU) Oct, 2007

[23]. Najiya Abdulrahiman, and Ranju K.V, "Text Dependent Speaker Recognition", *International Journal of Electrical, Electronics and Data Communication*, Volume- 1, Issue- 7, Sep-2013

[24]. Academia. [Online]. Available:
https://www.academia.edu/15548548/Speaker_Recognition, accessed on Feb 20, 2018

[25]. Ashish Kumar Panda and Amit Kumar Sahoo, "Study of Speaker Recognition Systems", National Institute Of Technology, Rourkela, 2007

[26]. Smitha Gangisetty, "Text-independent Speaker Recognition", Morgantown, West Virginia, 2005

[27]. Arun Rajsekhar. G, "Real Time Speaker Recognition using MFCC and VQ", National Institute of Technology Rourkela – 769008. 2008

[28]. Gish, H., and Schmidt, M., "Text-independent speaker identification," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18-32, Oct. 1994.

[29]. Kenneth S. Bordens and Bruce B. Abbott, "*Research Design and Methods A Process Approach*" EIGHTH EDITION, Indiana University—Purdue University Fort Wayne, 2011

[30]. Theodore Philip King Ernst, “Speaker Dependent Voice Recognition with Word-Tense Association and Part-Of-Speech Tagging”, Embry-Riddle Aeronautical University Daytona Beach, Florida, Dec, 2014

[31]. Sandeep Kaur, “A HMM Integrated SVM Model for Hindi Speech Recognition”, *Indian Journal of Science and Technology*, Vol 9(47), DOI: 10.17485/ijst/2016/v9i47/101744, Dec, 2016

[32]. Li Zhu,Qing Yang, “Speaker Recognition System Based on weighted feature parameter, *sciVerse sciencedirect*, 2012

[33]. Sanaa Hamid and Heamed Mohammed, “Speaker Identification System Using Wavelets and Neural Networks”, University of Khartoum, Jun, 2009

[34]. Joe Tebelskis, “Speech Recognition using Neural Networks”, School of Computer Science Carnegie Mellon University Pittsburgh, Pennsylvania, May, 1995

[35]. Tushar Ranjan Sahoo and Sabyasachi Patra, “Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification”, *I.J. Image, Graphics and Signal Processing*, vol. 6, page 27-35, May, 2014

[36]. R.M.Sneha, and K.L.Hemalatha, “Implementation of MFCC Extraction Architecture and DTW Technique in Speech Recognition System”, *International Journal of Emerging Trends in Science and Technology*, Vol.03,no. 05 , Pages 753-757, May, 2016.

[37]. Furui Sadaoki, "An overview of speaker recognition technology", 1-10, In *ASRIV*- April, 1994.

[38]. Articulator [online], available at:
https://en.wikipedia.org/wiki/Articulator, date accessed on Feb 20, 2018

[39]. Biological neural network [online], available at:
https://en.wikipedia.org/wiki/Biological_neural_network, date accessed on Feb, 2018

[40]. Ravi. B, “Speaker Dependent Continuous Speech Recognition of Hindi Language”, Department of Electrical Engineering, Indian Institute of Technology, Roorkee -247 667, June, 2006

[41]. Artificial neural network [online], available at:
https://en.wikipedia.org/wiki/Artificial_neural_network, date accessed on Feb 20, 2018

[42]. K.Sreenivasa Rao and Sourjya Sarkar, “Robust Speaker Recognition in Noisy Environments”, Springer Cham Heidelberg New York Dordrecht London, 2014

[43]. J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. Bonastre, and D.Matrouf, "Forensic Speaker Recognition," *IEEE Signal processing Magazine*, March, 2009.

[44]. J. P. Campbell, "Speaker Recognition for Forensic Applications," Lincoln Laboratory, Massachusetts Institute of Technology, 2014.

[45]. Wolelaw Berehane, “Amharic Text-Prompted Speaker Verification Model”, Master Thesis, Department of Computer Science, Addis Ababa University, July, 2011.

[46]. Sabyasachi Patra, “Robust Speaker Identification System”, Super Computer Education and Research Centre, Indian Institute of Science, BANGALORE 560 012, Dec, 2007

[47]. Karim Taam and Estelle Cherrier, “Speaker Recognition for Mobile user Authentication: an Android Solution”, *Research Gate,* Sep, 2013

[48]. VOICEBOX is a MATLAB toolbox for speech processing, [online]. Available at http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, date accessed on Feb 22, 2018

Appendix A: The Tigrigna character set elements

	ä	u	i	a	e	(ə)	O	wä	wi	wa	we	wə
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ					
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ					
ḥ	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ					
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ					
ś	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ					
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ					
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ					
š	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ					
ḳ	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቈ	ቊ	ቋ	ቌ	ቍ
ḳʰ	ቐ	ቑ	ቒ	ቓ	ቔ	ቕ	ቖ	ቘ	ቚ	ቛ	ቜ	ቝ
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ					
v	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ					
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ					
č	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ					
ẖ	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ	ኈ	ኊ	ኋ	ኌ	ኍ
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ					
ñ	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ					
ʾ	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ					

k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኰ	ኲ	ኳ	ኴ	ኵ
x	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ	ዀ	ዂ	ዃ	ዄ	ዅ
w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ					
ʿ	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ					
z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ					
ž	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ					
y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ					
d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ					
ǧ	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ					
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ	ጐ	ጒ	ጓ	ጔ	ጕ
ṭ	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ					
č̣	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ					
p̣	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ					
ṣ	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ					
ṣ́	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ					
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ					
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ					
	ä	**u**	**i**	**a**	**e**	**(ə)**	**O**	**wä**	**wi**	**wa**	**we**	**wə**

Appendix B: Source code of Speaker recognition from File

```
function varargout = Speaker_Recognition_ANN(varargin)
% SPEAKER_RECOGNITION_ANN MATLAB code for Speaker_Recognition_ANN.fig
%      SPEAKER_RECOGNITION_ANN, by itself, creates a new
SPEAKER_RECOGNITION_ANN or raises the existing
%      singleton*.
%
%      H = SPEAKER_RECOGNITION_ANN returns the handle to a new
SPEAKER_RECOGNITION_ANN or the handle to
%      the existing singleton*.
%
%      SPEAKER_RECOGNITION_ANN('CALLBACK',hObject,eventData,handles,...)
calls the local
%      function named CALLBACK in SPEAKER_RECOGNITION_ANN.M with the given
input arguments.
%
%      SPEAKER_RECOGNITION_ANN('Property','Value',...) creates a new
SPEAKER_RECOGNITION_ANN or raises the
%      existing singleton*.  Starting from the left, property value pairs are
%      applied to the GUI before Speaker_Recognition_ANN_OpeningFcn gets
called.  An
%      unrecognized property name or invalid value makes property application
%      stop.  All inputs are passed to Speaker_Recognition_ANN_OpeningFcn via
varargin.
%
%      *See GUI Options on GUIDE's Tools menu.  Choose "GUI allows only one
%      instance to run (singleton)".
%
% See also: GUIDE, GUIDATA, GUIHANDLES

% Edit the above text to modify the response to help Speaker_Recognition_ANN

% Last Modified by GUIDE v2.5 11-Apr-2018 20:28:27

% Begin initialization code - DO NOT EDIT
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                   'gui_Singleton',  gui_Singleton, ...
                   'gui_OpeningFcn', @Speaker_Recognition_ANN_OpeningFcn, ...
                   'gui_OutputFcn',  @Speaker_Recognition_ANN_OutputFcn, ...
                   'gui_LayoutFcn',  [] , ...
                   'gui_Callback',   []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end

if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end
% End initialization code - DO NOT EDIT

% --- Executes just before Speaker_Recognition_ANN is made visible.
```

```
function Speaker_Recognition_ANN_OpeningFcn(hObject, eventdata, handles,
varargin)
if (exist('SpeakersSoundDatabase.dat','file')==2)
    load('SpeakersSoundDatabase.dat','-mat');
    set(handles.SoundEdit,'String',sound_number);
end

handles.output = hObject;

% Update handles structure
guidata(hObject, handles);
function varargout = Speaker_Recognition_ANN_OutputFcn(hObject, eventdata,
handles)
varargout{1} = handles.output;
set(gcf,'CloseRequestFcn',@Speakerrecognition_closefcn)

function Speakerrecognition_closefcn(src,evnt)

 selection = questdlg('Do you want to Exit ?',...
      'Close Request Function',...
      'Yes','No','Yes');
   switch selection,
      case 'Yes',
         % we can add here what we want to save;
         delete(gcf)
      case 'No'
      return
   end
% --- Executes on button press in loadsoundfile.
function loadsoundfile_Callback(hObject, eventdata, handles)
clc;
global sound_file;
global filename;
global pathname;
global y;
global Fs;

[filename, pathname] = uigetfile('*.wav','select a wave sound file to load');
if pathname == 0
    errordlg('Error! sound file is not selected, please try again!');
    return;
end
sound_file = strcat(pathname,filename);
[y,Fs] = audioread(sound_file);
if ~exist('sound_file')
    errordlg('Error! path doesnot exist, please try again!');
    return;
end;
dt = 1/Fs;
t = 0:dt:(length(y)*dt)-dt;
plot(t,y); xlabel('Seconds'); ylabel('Amplitude');title('Time Domain');grid
on;
vfeatures  = mfcc(y,Fs); %% voice vfeatures MFCC vectors
disp('you  has  been selected sound . Now you can add it to database  ');
disp('or perform speaker recognition ');
```

```
msgbox('you seleted a sound succesffuly ,Now add it to database or Test
Speaker recognition','success');

% --- Executes on button press in recourdsound.
function recourdsound_Callback(hObject, eventdata, handles)

Microphone_Sound_Edit();
% --- Executes on button press in playsound.
function playsound_Callback(hObject, eventdata, handles)
global sound_file;
global pathname;
global filename;
sound_file = strcat(pathname,filename);
    if ~isempty(sound_file)
        [y,Fs]=audioread(sound_file);
        sound(y,Fs);
    else
        warndlg('no sound signal available','warning');
    end

% --- Executes on button press in AddsoundtoDB.
function AddsoundtoDB_Callback(hObject, eventdata, handles)
global  filename;
global  name;
global  pathname;
global  sound_number;
global  prmeter;
global  class_number;
global  max_class;
global  vfeatures;
global  features_data;
global  y;
global  Fs;
clc;
if exist('y')
    if (exist('SpeakersSoundDatabase.dat','file')==2)
        load('SpeakersSoundDatabase.dat','-mat');
        sound_number = sound_number+1;
        prompt={sprintf('%s','Please enter a positive integer for class
number <= ',num2str(max_class))};
        title='Class number';
        lines=1;
        def={'1'};
        answer=inputdlg(prompt,title,lines,def);
        prmeter=double(str2num(char(answer)));
        if size(prmeter,1)~=0
            class_number=prmeter(1);
            if
(class_number<=0)||(class_number>max_class)||(floor(class_number)~=class_numb
er)||(~isa(class_number,'double'))||(any(any(imag(class_number))))
                warndlg(sprintf('%s','Class number must be a positive integer
<= ',num2str(max_class)),' Warning ')
            else
                def={'name'};
                name=inputdlg('Enter Speakers Name ','Name',1,def);
```

```
                disp('Now the selected sound is going to MFCC feature extract
...wait');
                if class_number==max_class;
                    disp('this person sound class is not added to database so
far!');
                    max_class = class_number+1;
                    vfeatures  = mfcc(y,Fs);

                else
                    disp('this person sound class has  been added before to
database');
                    vfeatures  = mfcc(y,Fs);

                end
                features_data{features_size+1,1} = vfeatures;
                features_data{features_size+1,2} = class_number;
                features_data{features_size+1,3} = strcat(pathname,filename);
                features_data{features_size+1,4} = name;
                features_size= size(features_data,1);
                clc;

save('SpeakersSoundDatabase.dat','sound_number','max_class','features_data','
features_size','-append');
                msgbox(sprintf('%s','Database already exists: sound
succesfully added to class number ',num2str(class_number)),'Database
result','help');
                clc;
                disp('The added Sound to database Information is :.');
                mess = sprintf('%s','Location: ',strcat(pathname,filename));
                disp(mess);
                mess = sprintf('%s','Sound ID: ',num2str(class_number));
                disp(mess);
                a=cellstr(name);
                disp(strcat('Speaker name: ',a));
            end
        else
            warndlg(sprintf('%s','Class number must be a positive integer <=
',num2str(max_class)),' Warning ')
        end
    else
        sound_number = 1;
        max_class=1;
        prompt={sprintf('%s','Please enter a positive integer for class
number <= ',num2str(max_class))};
        title='Class number';
        lines=1;
        def={'1'};
        answer=inputdlg(prompt,title,lines,def);
        prmeter=double(str2num(char(answer)));
        if size(prmeter,1)~=0
            class_number=prmeter(1);
            if
(class_number<=0)||(class_number>max_class)||(floor(class_number)~=class_numb
er)||(~isa(class_number,'double'))||(any(any(imag(class_number))))
```

```
                warndlg(sprintf('%s','Class number must be a positive integer
<= ',num2str(max_class)),' Warning ')
            else
                def={'name'};
                name=inputdlg('Enter speakers Name ','Name',1,def);

                disp('Now the selected sound is going to MFCC feature extract
...please wait');

                max_class=2;
                vfeatures  = mfcc(y,Fs);

                disp('Completed.');
                features_data{1,1} = vfeatures;
                features_data{1,2} = class_number;
                features_data{1,3} = strcat(pathname,filename);
                features_data{1,4} = name;
                features_size      = size(features_data,1);

save('SpeakersSoundDatabase.dat','sound_number','max_class','features_data','
features_size');
                msgbox(sprintf('%s','Database was empty. Database has just
been created. Sound succesfully added to class number
',num2str(class_number)),'Database result','help');
                clc;
                disp(' The added Sound to database Information is: .');
                mess = sprintf('%s','Location: ',strcat(pathname,filename));
                disp(mess);
                mess = sprintf('%s','Sound ID: ',num2str(class_number));
                disp(mess);
                disp(strcat('Speakers Name: ',name));
            end
        else
            warndlg(sprintf('%s','Class number must be a positive integer <=
',num2str(max_class)),' Warning ')
        end
    end
else
    errordlg('No sound has been selected.','File Error');

end
if (exist('SpeakersSoundDatabase.dat','file')==2)
    load('SpeakersSoundDatabase.dat','-mat');
    set(handles.SoundEdit,'String',sound_number);
end

% --- Executes on button press in SpeakerrecognitionANN.
function SpeakerrecognitionANN_Callback(hObject, eventdata, handles)
global filename;
global  pathname;
global  RS;
global  vfeatures;
global y;
global Fs;
global  features_data;
```

```
if exist('y')
            if (exist('SpeakersSoundDatabase.dat','file')==2)
                load('SpeakersSoundDatabase.dat','-mat');
                disp('MFCC Features extraction for Speaker
recognition...please wait');
                vfeatures  = mfcc(y,Fs);
                mess = sprintf('%s','Input sound:
',strcat(pathname,filename));
                disp(mess);
                disp('---');
                disp('Training in progress...');
                tic;
                RS = sound_utterance_matching(vfeatures);
                toc;
                time=toc/3;
                disp('---');
                disp('Recognized the Speakers Sound ID is:');
                msgbox(sprintf('Speaker Recognition based on Artificial
Neural Networks Recognition Terminated'));
                disp(RS);
                set(handles.RecognitionIDedit,'String',RS);
                set(handles.timeedit,'String',time);
                for i=1:sound_number
                    if (RS==features_data{i,2})

set(handles.recognitionname,'String',features_data{i,4});
                        disp(strcat('Speakers Name: ',features_data{i,4}));
                    end
                end

            else
                warndlg('No sound processing is possible. Database is
empty.',' Warning ');
            end
else
    warndlg('Input sound must be selected.',' Warning ');
end

function RemoveDatabse_Callback(hObject, eventdata, handles)

 if (exist('SpeakersSoundDatabase.dat','file')==2)
     button = questdlg('Do you really want to remove the Database?');
     if strcmp(button,'Yes')
            load('SpeakersSoundDatabase.dat','-mat');

            delete('SpeakersSoundDatabase.dat');

clear('max_class','sound_number','features_data','features_size');
            msgbox('Database was succesfully removed from the current
directory.','Database removed','help');
            set(handles.SoundEdit,'String',0);
            set(handles.RecognitionIDedit,'String',0);
            set(handles.timeedit,'String',0);
            set(handles.recognitionname,'String','');
            clear all;
```

```
     end
 else
    msgbox('Database is empty.','Database result','help');
    clear all ;
  end

function DBinformation_Callback(hObject, eventdata, handles)
global  sound_number;
global  max_class;
global  features_data;

if (exist('SpeakersSoundDatabase.dat','file')==2)
    load('SpeakersSoundDatabase.dat','-mat');
    msgbox(sprintf('%s','Database has ',num2str(sound_number),' sound(s).
There are',num2str(max_class-1),' class(es). '),'Database result','help');
    disp('Sounds present in database:');
    disp('---');
    for i=1:sound_number
        mess = sprintf('%s','Location: ',features_data{i,3});
        disp(mess);
        mess = sprintf('%s','Sound ID: ',num2str(features_data{i,2}));
        disp(mess);
        disp(strcat('Speakers Name: ',features_data{i,4}));
        disp('---');
    end
else
    msgbox('Database is empty.','Database result','help');

end
function SoundEdit_Callback(hObject, eventdata, handles)

function SoundEdit_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function RecognitionIDedit_Callback(hObject, eventdata, handles)
function RecognitionIDedit_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
function timeedit_Callback(hObject, eventdata, handles)
function timeedit_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
% --------------------------------------------------------------------
function plot_Callback(hObject, eventdata, handles)
% --------------------------------------------------------------------
function about_Callback(hObject, eventdata, handles)

About();
% --------------------------------------------------------------------
```

```
function Exit_Callback(hObject, eventdata, handles)

selection = questdlg('Do you want to Exit?',...
      'Close Request Function',...
      'Yes','No','Yes');
   switch selection,
      case 'Yes',
        % we can add what we save befor exit
         delete(gcf)
      case 'No'
      return
   end
% --------------------------------------------------------------------
function timedomainesignal_Callback(hObject, eventdata, handles)
global Fs;
  global y;
  global sn;
sn= str2num(get(handles.SoundEdit,'string'));
if sn==0
    warndlg('no sound selected ','warning');
    return;
else
   dt = 1/Fs;
   t = 0:dt:(length(y)*dt)-dt;
   plot(t,y); xlabel('Time(Seconds)'); ylabel('Amplitude');grid on;
end
% --------------------------------------------------------------------
function periodogramplot_Callback(hObject, eventdata, handles)
)
global y;
global Fs;
  global sn;
 sn= str2num(get(handles.SoundEdit,'string'));
if sn==0
    warndlg('no sound selected ','warning');
    return;
else
plot(psd(spectrum.periodogram,y,'Fs',Fs,'NFFT',length(y)));
end
% --------------------------------------------------------------------
function specto_Callback(hObject, eventdata, handles)
global y;
global Fs;
global sn;
 sn= str2num(get(handles.SoundEdit,'string'));
if sn==0
    warndlg('no sound selected ','warning');
    return;
else
    figure
specgram(y,256,Fs);
end
% --------------------------------------------------------------------
function voicevector_Callback(hObject, eventdata, handles)
global y;
global Fs;
```

```
global sn;
sn= str2num(get(handles.SoundEdit,'string'));
if sn==0
    warndlg('no sound selected ','warning');
    return;
else
vfeatures  = mfcc(y,Fs);
plot(vfeatures);title('Voice Features');grid on
end
function recognitionname_Callback(hObject, eventdata, handles)
function recognitionname_CreateFcn(hObject, eventdata, handles)
if ispc && isequal(get(hObject,'BackgroundColor'),
get(0,'defaultUicontrolBackgroundColor'))
    set(hObject,'BackgroundColor','white');
end
```

Appendix C: Source code of FFANN model

```
function [net] = createfeedforwardnn(P,T)
alphabet = P;
targets  = T;
net=patternnet(25,'trainlm');
net.layers{1}.transferFcn = 'tansig';
net.layers{2}.transferFcn = 'tansig';
net.LW{1,1}          = net.LW{1,1}*0.01;
net.b{1}             = net.b{1}*0.01;
net.performFcn        = 'mse';
net.trainParam.goal   = 0;
net.trainParam.show   = 25;
net.trainParam.epochs = 1000;
net.trainParam.mc     = 0.95;
net.trainParam.lr=0.01;
net.trainParam.showWindow=0;
net.divideFcn='dividerand';
net.divideMode = 'sample';  % Divide up every sample
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;
 net.input.processFcns = {'removeconstantrows','mapminmax'};
 net.output.processFcns = {'removeconstantrows','mapminmax'};
net.trainFcn = 'trainlm';
P                    = alphabet;
T                    = targets;
net = init(net);
[net] = train(net,P,T);
   y=net(P);
figure, plotconfusion(T,y)

view(net);
```