Amharic Speech Recognition using Power Spectrum Density Estimation and Discrete Wavelet Transform

Daniel Lemma Daba

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in partial fulfillment of the Requirement for the Degree of Master's in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

June, 2022

Adama, Ethiopia

Amharic Speech Recognition using Power Spectrum Density Estimation and Discrete Wavelet Transform

Daniel Lemma Daba

Advisor: Dr. Tilahun Melak

A Thesis Submitted to the Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in partial fulfillment of the Requirement for the Degree of Master's in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

June, 2022

Adama, Ethiopia

DECLARATION

I hereby declare that this Master Thesis entitled ‘‘**Amharic Speech Recognition using Power Spectrum Density Estimation and Discrete Wavelet Transform**’’ is my original work. That is, it has not been submitted for the award of any academic degree, diploma or certificate in any other university. All sources of materials that are used for this thesis have been duly acknowledged through citation.

Daniel Lemma Daba		
Name of the student	Signature	Date

RECOMMENDATION

I/we, the advisor(s) of this thesis, hereby certify that I/we have read the revised version of the thesis entitled ‘‘**Amharic Speech Recognition using Power Spectrum Density Estimation and Discrete Wavelet Transform**’’ prepared under my guidance by Daniel Lemma Daba submitted in partial fulfillment of the requirements for the degree of Mater’s of Science in Computer Science and Engineering. Therefore, I recommend the submission of revised version of the thesis to the department following the applicable procedures.

Dr. Tilahun Melak		
Major Advisor	Signature	Date
Co-advisor	Signature	Date

APPROVAL PAGE

I/we, the advisors of the thesis entitled ‘‘**Amharic Speech Recognition using Power Spectrum Density Estimation and Discrete Wavelet Transform**’’ and developed by Daniel Lemma Daba, hereby certify that the recommendation and suggestions made by the board of examiners are appropriately incorporated into the final version of the thesis.

Dr. Tilahun Melak Major Advisor	Signature	Date
Co-advisor	Signature	Date

APPROVAL OF BOARD OF EXAMINERS

We, the undersigned, members of the Board of Examiners of the thesis by **Daniel Lemma Daba** have read and evaluated the thesis entitled "**Amharic Speech Recognition using Power Spectrum Density Estimation and Discrete Wavelet Transform**" and examined the candidate during open defense. This is, therefore, to certify that the thesis is accepted for partial fulfillment of the requirement of the degree of Master of Science in Computer Science and Engineering.

Chairperson	Signature	Date
Internal Examiner	Signature	Date
External Examiner	Signature	Date

Finally, approval and acceptance of the thesis is contingent upon submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC).

Department Head	Signature	Date
School Dean	Signature	Date
Office of Postgraduate Studies, Dean	Signature	Date

ACKNOWLEDGMENT

First of all, I would like to thank Holy God with his Holy Virgin Mother Mary for everything they have done for me and for their protection.

I thank my family for their support they have given me in any means they possibly can. I also should like to thank my colleagues for their support in searching for the title and advises they gave me when things get difficult at the time of doing this thesis work.

I should give special thanks to my advisor Dr. Tilahun for his valuable time, advice and support in the process of this thesis work. He encourages me in every dimension of he possibly can. He has removed all my worries from the beginning to the end of this thesis work by reminding me of what my research requires.

TABLE OF CONTENT

LIST OF TABLES

TABLE **PAGE**

LIST OF FIGURES

FIGURE **PAGE NO.**

LIST OF ACRONYMS AND ABBREVIATIONS

AFTI	Advanced Fighter Technology Integration
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
ASRS	Automatic Speech Recognition System
CTC	Connectionist Temporal Classification
DARPA	Defense Advanced Research Projects Agency
DFT	Discrete Fourier Transform
DL	Deep Learning
DNN	Deep Neural Network
DRNN	Deep Recurrent Neural Network
DWT	Discrete Wavelet Transform
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
GUI	Graphical User Interface
HMM	Hidden Markov Model
IARPA	Intelligence Advanced Research Projects Activity
IEEE	Institute of Electrical and Electronics Engineers
MFCC	Mel frequency cepstral Coefficients
LPC	Linear Predictive Analysis
LSTM	Long short-term memory
OOV	Out Of Vocabulary
PLP	Perceptual linear predictive
PSDE	Power Spectrum Density Estimation
RNN	Recurrent neural networks
SRS	Speech Recognition System
TDNN	Time Delay Neural Networks
VQ	Vector Quantization
WER	Word Error Rate
WT	Wavelet Transform

ABSTRACT

This thesis work explored the possibility of developing Amharic Automatic Speech Recognition System. Towards this end, literature was reviewed on speech recognition, application of speech recognition, Amharic speech recognition. To develop and test the required Amharic speech recognition system, speech data were recorded, collected, transcribe to text format. Amharic phonetics and phonologies are widely differing from English language. This thesis work focuses on Automatic Speech Recognition for Amharic. Speech recognition is the ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format. Rudimentary speech recognition software has a limited vocabulary of words and phrases, and it may only identify these if they are spoken very clearly. English remains to be the most widely spoken languages of the world. While Automatic Speech Recognition research work prevails for English. However, there are few researches on speech technology in Ethiopian languages in general and Amharic in particular. Researches done for Amharic Speech Recognition are domain specific. These include Speech Translation for tourism areas, word recognition. The aim of this work is to enhance continues speech recognition for Amharic speech recognition by using a parametric method of Power Spectrum Density Estimation and Discrete Wavelet Transform. Speech dataset for Amharic language is a fundamental requirement for any development on Amharic Automatic Speech Recognition. Power Spectral Estimation method is to obtain an approximate estimation of the power spectral density of a given real random process. Discrete Wavelet Transform is applied to extract the approximation coefficients. Power spectral density is then applied to estimate the power spectrum density. The learning process is done by using Recurrent Neural Network (RNN) which is a class of Artificial Neural Networks that can process a sequence of inputs in deep learning and retain its state while processing the next sequence of inputs. For the training and evaluation purpose, we use 9340 audio files with their transcription which is 20hr audio file or around 5GB file is used. At the beginning of the training the Word Error Rate was high (19) and at the end of the training, word error rate is decreased to 7.02 and training loss decrease highly from the starting epoch to the end of the epoch which shows that the more we train the more decrease in Word Error Rate.

Keywords: *Automatic Speech Recognition, Discrete Wavelet Transform, Power Spectrum Density, Deep Learning, Recurrent Neural Network.*

CHAPTER ONE

INTRODUCTION

1.1 Background

Automatic speech recognition (ASR) is the process and the related technology for converting the speech signal into its corresponding sequence of words or sentence or other linguistic units by means of algorithms implemented in a device, a computer, or computer clusters or other type of machine. The speech signal contains not only the intended message, but also the characteristics of the utterance speaker and the language of communication. Basically, it means talking to your computer, and having it correctly recognize what you are saying. It is used to identify the words a person has spoken or to authenticate the identity of the person speaking into the system. Natural form of communication in humans being is speech.

For common people speech is just the sound waves coming out of the human mouth and perceived through ears (Kiran.R, Nivedha.K, Subha.T and Pavithra Devis, 2017). But there is complex mechanism behind its production. The study of human speech production and perception mechanism is important for the development of devices for hearing aids, speech recognition, speech enhancement, speech simulation, speech modeling etc. Nowadays the communication is not only between humans, it is also between humans and machines. Recently, advances in Automatic Speech Recognition (ASR) techniques have enabled machines to effectively communicate with humans. This type of communication is called speech recognition. Speech recognition is to solve so many problems that are facing people's day today activities by helping them to use their voice. However, this technological advancement is still far from Amharic. Amharic, like other languages advanced in speech recognition, should also be enriched with speech recognition. Researches done for Amharic Speech Recognition are domain specific. These include Speech Translation for tourism areas, word recognition. The aim of this work is to explore various possibilities for developing large vocabulary speaker independent continues speech recognition for Amharic speech recognition system by using more data, different speakers and different feature extraction methods to get better result. To achieve these objectives, we use a variety of methods, tools and techniques. Among these, using power spectrum density estimation and discrete wavelet transform are the basic ones. But before using these methods, data set collection is the fires activity. After the necessary data set is collected, we will use the above mentioned and other methods to achieve our objectives.

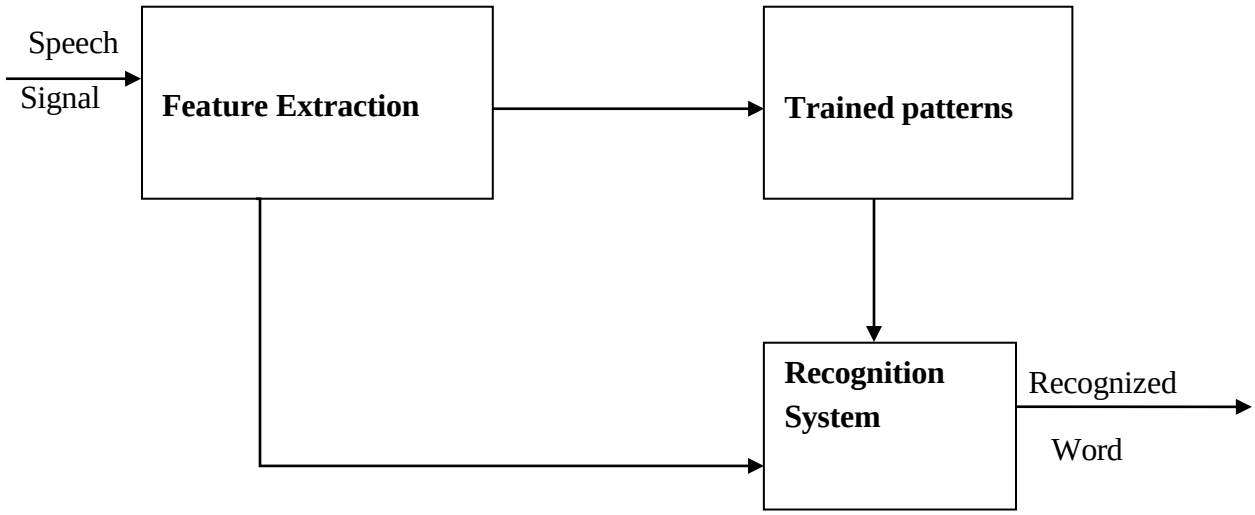

Figure 1.1 Speech Recognition System Block Diagram

Speech recognizers aim to extract the lexical information from the speech signal independently of the speaker by reducing the inter-speaker variability (Kiran.R, Nivedha.K, Subha.T and Pavithra Devis, 2017). Current speech recognition systems are not ideal compared to a human's capacity in understanding speech (1). A number of problems are not solved yet. The main problem of speech recognition lies in the variation of speech. A message can be spoken quite differently by varying the speaking rate, the individual sounds, the speaker, the gender of the speaker (Maazouzi, A. Laaroussi, N. Aqili, M. Raji and A. Hammouch LRGE, ENSET, Mohammed V, 2016). There are several approaches to implement automatic speech recognition (Maazouzi, A. Laaroussi, N. Aqili, M. Raji and A. Hammouch LRGE, ENSET, Mohammed V, 2016) the statistical approach by using Hidden Markov Models and the connectionist approach by using Neural Networks. Also Dynamic Time Warping (DTW) algorithm based on dynamic programming is abundantly used for the automatic speech recognition (Arkady Produs, Kateryna Kukharicheva, 2017). The performance of a speech recognition system is measurable and the most widely used method for measuring the performance is calculating the recognition accuracy. Speech signals involve rich information, including linguistic content, speaker trait, emotion, channel and background noise, etc. Researchers have worked for several decades to decode this information, leading to a multitude of speech information processing tasks, including automatic speech recognition (ASR). There are two important phases that describe a speech recognition system: features extraction and recognition (Maazouzi, A. Laaroussi, N. Aqili, M.

Raji and A. Hammouch LRGE, ENSET, Mohammed V, 2016).

A vector is extracted from the speech signal of each spoken word during the first phase of speech recognition. In this phase, it converts the speech signal into a sequence of acoustic feature vectors to extract a number of features from the signal that have a maximum of information relevant for classification (Basem H. A. Ahmed and Ayman S.Ghabayen, 2017). The number of features extracted from the waveform signal is commonly much lower than the number of signal samples, thus reducing the amount of data (Basem H. A. Ahmed and Ayman S.Ghabayen, 2017). The training vectors (Maazouzi, A. Laaroussi, N. Aqili, M. Raji and A. Hammouch LRGE, ENSET, Mohammed V, 2016) extract the features to distinguish various classes of words. Each training vector helps to construct model for word and the resulted vectors are arranged in a database to be used in the next step. The user pronounces any word for which the system was trained in the second phase of speech recognition. There are many features extraction techniques, the most frequent are linear predictive Analysis (LPC), perceptual linear predictive (PLP) coefficients, Mel frequency cepstral coefficients (MFCC), power spectrum analysis such as Fast Fourier Transform (FFT) etc. (Maazouzi, A. Laaroussi, N. Aqili, M. Raji and A. Hammouch LRGE, ENSET, Mohammed V, 2016). Speech Recognition offers greater freedom to employ the physically handicapped in several applications like manufacturing processes, medicine, military and telephone network.

1.2 Statement of the Problem

Amharic, which belongs to the Semitic language family, is the official language of Federal Government Ethiopia and the most commonly learned second language throughout the country. In this family, Amharic stands second in its number of speakers after Arabic (Abraham Woubie Zewoudie, 2011). The ultimate goal of research in speech technology is the development of human-machine conversational system that communicates with any one, about anything, on any topic and in any situation. To achieve this goal, researches have been conducted for several years and various systems are developed for various languages. However, there are few researches on speech technology in Ethiopian languages in general and Amharic in particular. These researches were domain specific, isolated word recognition, connected word recognition, Translation for tourism areas, speaker dependent, speaker independent with small amount of data, command based word recognition for the domain they are done for. Automatic speech recognition systems work with a pre-defined lexicon, i.e., the number of distinct words it contains, which is an important parameter for an Automatic Speech Recognition system. For various reasons, words in Amharic are pronounced differently and varied from one speaker to another and from one situation to another.

Especially at real-time there will be unknown word display because of environmental noise, speaker variability and for other reasons. This is termed as pronunciation variation and is one of the major factors that degrade the performance of any Automatic Speech Recognition systems. this thesis work focuses on solving the pronunciation variation problem by adding multiple pronunciations to words that have more than one pronunciation in the lexicon, by using more speech data and acquiring the actual pronunciation for each word in the lexicon and trying to show Amharic can be used for speaker independent continues speech recognition because previous works domain specific like tourism word translation with word recognition accuracy of 80.6 And 49.3 recognition accuracy of sentence.

1.3 Research Question

This thesis work tried to answer the following basic questions regarding to the Amharic Automatic Speech Recognition and challenges that affect the ASR

I. How much speech corpora are needed to develop ASR?
II. What model and technique is needed?
III. What are the challenges to develop continuous, speaker independent speech recognition?
IV. What are the factors affecting quality and performance of ASR?

1.4 Objectives

1.4.1 General Objectives

The general objective of this work is to explore the possibility of developing Amharic continuous Speech Recognition System using Power Spectrum Density Estimation and Discrete Wavelet Transform.

1.4.2Specific Objectives

The specific objectives are

I. To develop speech corpus that can be used for training and testing purpose
II. To design lexical and acoustic model
III. To incorporate multiple pronunciations sample set into Amharic lexicon
IV. To implement speaker independent and continuous speech recognition
V. To test and analyze the performance of the ASR System
VI. To analyze the results and give conclusion and forward recommendations

1.5 Significance of the Study

Speech recognition will be a key part of the future of communication. Communicating with technological devices via voice has become so popular. Many businesses are adopting this new way of working, whether to improve their own internal processes or upgrade their customer service systems. Despite this, speech recognition is still relatively new. But most of the communications are held on in English and other International languages. Developing Amharic ASR system is crucial to help People with disabilities, for the area of health care, military, Telephony and other domains.

1.6 Expected Outcome

The following points would be the outcome of the study: -

I. Prepared Dataset for Amharic ASRS.
II. Sample Amharic ASRS with GUI.
III. Experimentally tested results.

1.7 Delimitations and Scope

1.7.1 Scope

The scope of this thesis is limited to modeling Amharic Speech Recognition System with minimum recognition errors due to pronunciation variation and thus enhances the performance of large vocabulary, Continuous, speaker independent Amharic speech recognition system. The best result for Amharic Speech Recognition is 80.6% and 49.3% at word and sentence level respectively. The aim of this thesis is to implement Continuous, speaker independent Amharic speech recognition system with minimum error.

1.7.2 Limitation

The aforementioned general and specific objectives may not be fully implemented or there may be lack of accuracy for some of the following reasons**.**

I. Noise-robustness: - Speech is uttered in an environment of sounds, a radio playing somewhere down the corridor, another human speaker in the background etc. This is unwanted information in the speech signal. This may degrade the accuracy of the ASRS.
II. Speaker/accent variability: - All speakers have their special voices, due to their unique physical body and personality.
III. Amount of dataset: - less amount of data set can cause OOV error.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

In this chapter, we have presented related works on Automatic Speech Recognition Systems done in other languages and reviewed the works done in Amharic Automatic Speech Recognition System previously. Most of the related works reviewed are used as an input to this work. As we have reviewed, most of researched done on Amharic Automatic Speech Recognition System are domain specific. During doing this work, we can't found a research done for speaker independent continuous Automatic Speech Recognition System for Amharic.

2.2 Theoretical and Historical background of Automatic Speech Recognition

2.2.1 Automatic Speech Recognition

Speech recognition is an interdisciplinary subfield of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text or word by computers. It is also known as automatic speech recognition (ASR), computer speech recognition or speech to text (STT) (Ankita Chugh, Poonam Rana, Suraj Rana, 2014). It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields. Most of speech recognition systems require "training" (also called "enrollment") where an individual speaker reads text or word or isolated vocabulary into the system. The system analyzes the person's specific voice characteristics and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy and processing speed. Systems that do not use training are called "speaker independent" systems. Systems that use training are called "speaker dependent". Speech recognition applications include voice user interfaces such as voice dialing (e.g. "call home"), call routing (e.g. "I would like to make a collect call"), domestic appliance control, search key words (e.g. find a podcast where particular words were spoken), simple data entry (e.g., entering a credit card number), preparation of structured documents (e.g. a radiology report), determining speaker characteristics, speech-to-text processing (e.g., word processors or emails), and controlling industrial appliances. The term voice recognition or speaker identification refers to identifying the speaker or who is speaking, rather than what they are saying. Recognizing the speaker can simplify the task of translating speech

or the voice of the speaker in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of a speaker as part of a security process (Rainer Martin, 2019). From the technology perspective, speech recognition has a long history with several waves of major innovations. Most recently, the field has benefited from advances in deep learning and big data. The advances are evidenced not only by the surge of academic papers published in the field, but more importantly by the worldwide industry adoption of a variety of deep learning methods in designing and deploying speech recognition systems.

Amharic Language

Amharic is used as a first language for many communities and as a lingua franca by others to serve as the official Ethiopian language for business and administrative duties in all cities and towns. According to current data, nearly 22 million people speak Amharic as their first language in Ethiopia and it has over 4 million second-language speakers within the country and a further 3 million around the world. Amharic in Ethiopia is used in commerce, government, media and national education. The majority of the speakers of Amharic found in Ethiopia, but there are also speakers in a number of other countries, particularly Israel, Eritrea, Canada, the USA and Sweden. It has five dialectical variations across different parts of the country (Abraham Woubie Zewoudie, 2011). These include dialects like Addis Ababa, Gojam, Gonder, Wollo and Menz (Vimal krishinan VR, Athulya Jayakumar, Babu Anto.P, 2018). It belongs to the Semitic language family and is characterized by a quite homogeneous phonology distinguishing between 234 distinct Consonant-Vowel (CV) syllables. Amharic has five dialectical variations spoken in different Amharic regions: Addis Ababa, Gojjam, Gonder, Wollo, and Menz. The dialect of Addis Ababa has emerged as the standard dialect and has wide currency across all Amharic-speaking communities (Tanja Schultz the 9th International symposium on chines spoken language processing, 2014)

Unlike other Semitic languages, such as Arabic and Hebrew, Amharic script uses a grapheme based writing system called Fidel which is written and read from left to right. Modern Amharic has inherited its writing system from Ge'ez /gəʾəzə/, which is still the classical and ecclesiastical language of Ethiopia an Amharic character represents a consonant vowel (CV) sequence and the basic shape of each character is determined by the consonant, which is modified for the vowel. Automatic speech recognition (ASR) is the process of converting an acoustic signal, captured by microphone or a telephone, to a set of text and is a broad area of research. Automatic speech recognition uses machine-

based process of decoding and transcribing oral speech and related technology for converting speech signals into a sequence of words or other linguistic units by means of an algorithm implemented as a computer program (Kiran.R, Nivedha.K, Subha.T and Pavithra Devis, 2017). A set of 38 phones, seven vowels and 31 consonants, makes up the complete inventory of sounds for the Amharic language. Consonants are generally classified as stops, fricatives, nasals, liquids and semi vowels. Table 2.1 shows the phonetic representation of Amharic consonants as to their manner of articulation, voicing and place of articulation. Some of the Amharic consonants have similar phonetic transcriptions like English. These include ብ [b], ድ [d], ፍ [f], ግ [g], ሀ [h], ክ [k], ል [l], ም [m], ን [n], ፕ [p], ር [r], ስ [s], ት [t], ቭ ው[w], [y] and ዝ[z]. They correspond to English

consonants b, d, f, g, h, k, l, m, n, p, r, s, t, v, w, y, and z. There are also sounds that have similar sound as English sounds, but are represented using different symbols. These include ች[ch], ኝ [nx], ሽ[sx] and ዥ[zx]. Sounds that are the characteristics of Amharic but not found in English are ጵ[px], ጥ[tx], ፅ[xx], ጭ[cx] and ቅ[q] (Solomon, 2006). In Amharic, all consonants except ሀ/h/ and ዕ/ax/ may occur in either a geminated or a non-geminated form. The germination of Amharic is not shown in the orthography. Germination is the lengthening in time of the consonants (Vimal krishinan VR, Athulya Jayakumar, Babu Anto.P, 2018).

Table 2.1 Categories of Amharic Consonants

		Labials		Alveolar		Palatals		Velars		Labio-Velar		Glottals	
Stops	Voiceless	p	ፕ	t	ት			k	ክ	kwa	ኳ	ax	ዕ
	Voiced	b	ብ	d	ድ			g	ግ	gwa	ጓ		
	Glottalized	px	ጵ	tx	ጥ			q	ቅ	qwa	ቋ		
Fricatives	Voiceless	f	ፍ	s	ስ	sx	ሽ					h	ሀ
	Voiced			z	ዝ	zx	ዥ						
	Glottalized			xx	ጽ							hwa	ኋ
Affricatives	Voiceless					c	ች						
	Voiced					j	ጅ						
	Glottalized					cx	ጭ						
Nasals	Voiced	m	ም	n	ን	nx	ኝ						
Liquids	Voiced			l	ል								
	Voiced			r	ር								
Glides		w	ው			y	ይ						

The seven vowels, along with their representation in Ge'ez characters, are shown in terms of their place of articulation in Table 2.2. In addition to the five vowels common among many languages, Amharic has two central vowels, /e/ and /ix/, the latter with a mainly epenthetic function. The epenthetic vowel /ix/ plays a key role in syllabification.

Table 2.2: Categories of Amharic Vowels

	Front	Central	Back
High	ኢ [ii]	እ [ix]	ኡ [u]
Mid	ኤ [ie]	ኧ [e]	ኦ [o]
Low		አ [a]	

Amharic Writing System

Amharic uses the Ge'ez (or Ethiopic) writing system which is originated with the ancient Ge'ez

language. In the writing system, even if the symbols are consonant-based, they also contain an obligatory vowel marking. Most of the symbols represent consonant vowel combinations. Each Amharic consonant is associated with seven characters (referred to as ‘‘orders’’) for the seven vowels of the language. It is the sixth-order character that is the special symbol, representing the plain consonant. The basic pattern for each consonant is shown in Table. 2.3.

Table 2.3: Amharic syllabic structure with example for consonant

1st order	2nd order	3rd order	4th order	5th order	6th order	7th order
m[e]	m[u]	m[ii]	m[a]	m[ie]	m	m[o]
መ	ሙ	ሚ	ማ	ሜ	ም	ሞ

2.2.2 History of Automatic Speech Recognition System

Pre-1970

1952 – Three Bell Labs researchers, Stephen Balashek, R. Biddulph, and K. H. Davis built a system called "Audrey" for single-speaker digit recognition. Their system located the formants in the power spectrum of each utterance (Pavel Krbec, 2018). In the year of 1960 – Gunnar Fant developed and published the source-filter model of speech production. Latter on the year of 1962 IBM demonstrated its 16-word "Shoebox" machine's speech recognition capability at the 1962 World's Fair (Pavel Krbec, 2018). The new speech coding method which is Linear predictive coding (LPC), was first proposed by Fumitada Itakura of Nagoya University and Shuzo Saito of Nippon Telegraph and Telephone (NTT), while working on speech recognition on the year 1966 (Jungsuk Kim and Wonyong Sung., 2012). In 1969 Funding at Bell Labs dried up for several years when, in 1969, the influential John Pierce wrote an open letter that was critical of and defunded speech recognition research. This defunding lasted until Pierce retired and James L. Flanagan took over.

In the late 1960s, continuous speech recognition was introduced by Raj Reddy at Stanford University. Previous systems required users to pause after each word. Reddy’s system issued spoken commands for playing game chess.

Around this time Soviet researchers invented the dynamic time warping (DTW) algorithm and used it to create a recognizer capable of operating on a 200-word vocabulary (Kim-Y. E. Wong, 2014). DTW processed speech by dividing it into short frames, e.g. 10ms segments, and processing each frame as a single unit. Although DTW would be superseded by later algorithms, the technique carried on. Achieving speaker independence remained unsolved at this time period (S. Ranjan, C. Yu, C. Zhang,

F. Kelly, J. H. L. Hansen, 2016, Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk, 2017).

From 1970 to 1990

Defense Advanced Research Projects Agency (DARPA) funded five years for Speech Understanding Research in the year 1971. The Research was, Speech recognition research seeking a minimum vocabulary size of 1,000 words. They thought speech understanding would be key to making progress in speech recognition. In the year, 1976 The first ICASSP was held in Philadelphia, which since then has been a major venue for the publication of research on speech recognition (Jungsuk Kim and Wonyong Sung,2012). A decade later, at CMU, Raj Reddy's students James Baker and Janet M. Baker began using the Hidden Markov Model (HMM) for speech recognition. The use of HMMs allowed researchers to combine different sources of knowledge, such as acoustics, language, and syntax, in a unified probabilistic model (Martha Yifru Tachbelie, Solomon Teferra Abate, Wolfgang Menzel, 2017). By the mid-1980s IBM's Fred Jelinek's team created a voice activated typewriter called Tangora, which could handle a 20,000-word vocabulary Jelinek's statistical approach put less emphasis on emulating the way the human brain processes and understands speech in favor of using statistical modeling techniques like HMMs. (Jelinek's group independently discovered the application of HMMs to speech.) This was controversial with linguists since HMMs are too simplistic to account for many common features of human languages. However, the HMM proved to be a highly useful way for modeling speech and replaced dynamic time warping to become the (Martha Yifru Tachbelie, Solomon Teferra Abate, Wolfgang Menzel, 2017, Jungsuk Kim and Wonyong Sung,2012)

Practical speech recognition

The 1980s also saw the introduction of the n-gram language model. In 1987 the back-off model allowed language models to use multiple length n-grams, and CSELT used HMM to recognize languages (both in software and in hardware specialized processors, e.g. RIPAC). Much of the progress in the field is owed to the rapidly increasing capabilities of computers. At the end of the DARPA program in 1976, the best computer available to researchers was the PDP-10 with 4 MB ram. It could take up to 100 minutes to decode just 30 seconds of speech (Jungsuk Kim and Wonyong Sung,2012).

Speech recognition products

In the year 1987 A recognizer from Kurzweil Applied Intelligence Dragon Dictate, a consumer product

released in 1990 AT&T deployed the Voice Recognition Call Processing service in 1992 to route telephone calls without the use of a human operator. The technology was developed by Lawrence Rabiner and others at Bell Labs. By this point, the vocabulary of the typical commercial speech recognition system was larger than the average human vocabulary (Pavel Krbec, 2018).

Speech Recognition in 2000s

In the 2000s DARPA sponsored two speech recognition programs: Effective Affordable Reusable Speech-to-Text (EARS) in 2002 and Global Autonomous Language Exploitation (GALE). Four teams participated in the EARS program: IBM, a team led by BBN with LIMSI and University of Pittsburgh, Cambridge University, and a team composed of ICSI, SRI and University of Washington. EARS funded the collection of the Switchboard telephone speech corpus containing 260 hours of recorded conversations from over 500 speakers. The GALE program focused on Arabic and Mandarin broadcast news speech. Google's first effort at speech recognition came in 2007 after hiring some researchers from Nuance. The first product was GOOG-411, a telephone based directory service (Pavel Krbec, 2018). The recordings from GOOG-411 produced valuable data that helped Google improve their recognition systems. Google Voice Search is now supported in over 30 languages. In the United States, the National Security Agency has made use of a type of speech recognition for keyword spotting since at least 2006. This technology allows analysts to search through large volumes of recorded conversations and isolate mentions of keywords.

Recordings can be indexed and analysts can run queries over the database to find conversations of interest. Some government research programs focused on intelligence applications of speech recognition, e.g. DARPA's EARS's program and IARPA's Babel program (Jungsuk Kim and Wonyong Sung, 2012). In the early 2000s, speech recognition was still dominated by traditional approaches such as Hidden Markov Models combined with feedforward Artificial Neural Networks (ANN). Today, however, many aspects of speech recognition have been taken over by a Deep Learning (DL) method called Long short-term memory (LSTM), a Recurrent Neural Network published (RNN) by Sepp Hochreiter & Jürgen Schmidhuber in 1997 (Pavel Krbec, 2018, Jungsuk Kim and Wonyong Sung, 2012). LSTM RNNs avoid the vanishing gradient problem and can learn "Very Deep Learning" tasks that require memories of events that happened thousands of discrete time steps ago, which is important for speech. Around 2007, LSTM trained by Connectionist Temporal Classification (CTC) started to outperform traditional speech recognition in certain applications (Pavel Krbec, 2018). In 2015, Google's speech recognition reportedly experienced a dramatic performance

jump of 49% through CTC-trained LSTM, which is now available through Google Voice to all smartphone users. The use of deep feedforward (non-recurrent) networks for acoustic modeling was introduced during later part of 2009 by Geoffrey Hinton and his students at University of Toronto and by Li Deng and colleagues at Microsoft Research, initially in the collaborative work between Microsoft and University of Toronto which was subsequently expanded to include IBM and Google. In contrast to the steady incremental improvements of the past few decades, the application of deep learning decreased word error rate by 30%. This innovation was quickly adopted across the field. Researchers have begun to use deep learning techniques for language modeling as well. In the long history of speech recognition, both shallow form and deep form of artificial neural networks had been explored for many years during 1980s, 1990s and a few years into the 2000s. But these methods never won over the non-uniform internal-handcrafting Gaussian mixture model/Hidden Markov model (GMM-HMM) technology based on generative models of speech trained discriminatively. A number of key difficulties had been methodologically analyzed in the 1990s, including gradient diminishing and weak temporal correlation structure in the neural predictive models. All these difficulties were in addition to the lack of big training data and big computing power in these early days. Most speech recognition researchers who understood such barriers hence subsequently moved away from neural nets to pursue generative modeling approaches until the recent resurgence of deep learning starting around 2009–2010 that had overcome all these difficulties (Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk,2017, Pavel Krbec, 2018).

Speech Recognition in 2010s

By early 2010s speech recognition, also called voice recognition was clearly differentiated from speaker recognition, and speaker independence was considered a major breakthrough. Until then, systems required a "training" period. A 1987 ad for a doll had carried the tagline "Finally, the doll that understands you." Despite the fact that it was described as "which children could train to respond to their voice". In 2017, Microsoft researchers reached a historical human parity milestone of transcribing conversational telephony speech on the widely benchmarked Switchboard task. Multiple deep learning models were used to optimize speech recognition accuracy. The speech recognition word error rate was reported to be as low as 4 professional human transcribers working together on the same benchmark, which was funded by IBM Watson speech team on the same task (Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk,2017, Pavel Krbec, 2018).

2.3 Types of Speech Recognition System

Speech Recognition Systems can be classified into different groups depending on the constraints imposed on the nature of the input speech, the speaker and the function of the speech recognition system. These constraints include: speaker dependence, type of utterance, size of the vocabulary, linguistic constraints, type of speech, Recognition Style and environment of use. We will describe some of the constraint as follows (Ankita Chugh, Poonam Rana, Suraj Rana, 2014, Vimal krishinan VR, Athulya Jayakumar, Babu Anto.P, 2018).

2.3.1 Speaker Dependent

Speech recognition system that can only recognize the speech or voice of users it is trained to understand. Speaker-dependent speech recognition systems are found in specialized use cases where there a limited number of words or phrases that need to be recognized with high accuracy, and processing time. The system learns the unique, individual characteristics of the speaker's voice through voice training. This type of system must be trained on a specific user before being able to recognize what has been said. This type of speech recognition system works well if there is only going to be specific user who will be speaking to the system. Speaker dependent speech recognition systems are generally able to recognize speech from a variety of contexts (words, phrases). New users must first ‘‘train’’ the system by speaking to it, so the computer can analyze the way in which the person talks (Ankita Chugh, Poonam Rana, Suraj Rana, 2014). Due to the fact that the acoustic variations among different speakers and difficulty to describe and model, Speaker Dependent Speech Recognition System usually perform much better than Speaker Independent Speech Recognition System. The drawback to this approach is that the system only responds accurately only to the individual who trained the recognition system. This is the most common approach employed in software for personal computers and fro security (Vimal krishinan VR, Athulya Jayakumar, Babu Anto.P, 2018).

2.3.2 Speaker Independent

A speech recognition system is said to be speaker independent if it can recognize speech of any and every speaker, such a system has learnt the characteristics of a large number of speakers. This type of speech recognition is a continuous speech recognition. A large amount of a user’s speech data (large amount of vocabulary) is necessary for training. Speaker-independent systems try to match the user’s voice to common voice patterns. This means that speaker-independent systems have an increased likelihood of errors and voice commands failing to be understood by the system, especially if the

user has a pronunciation or is not a native speaker of the language stated. Industrial requirements more often need speaker independent voice systems, such as the AT&T system used in the telephone systems (Kiran.R, Nivedha.K, Subha.T and Pavithra Devis,2017, Ankita Chugh, Poonam Rana, Suraj Rana, 2014).

2.4 Recognition Style

Speech recognition systems have another constraint regarding the style of speech they can recognize. They are three styles of speech: isolated, connected and continuous.

Isolated speech recognition systems can just handle words that are spoken separately or single word at a time. This is the most common speech recognition systems available today. The user must pause between each word or command spoken. This is the simplest form of recognition to perform because the end points are easier to find and the pronunciation of a word tends not affect others. Thus, because the occurrences of words are more consistent they are easier to recognize (Ankita Chugh, Poonam Rana, Suraj Rana, 2014).

Continuous is the natural conversational speech we are used to in everyday life. It is extremely difficult for a recognizer to shift through the text as the word tends to merge together. Continuous speech recognition systems are under continual development. To develop Continuous Speech Recognition System, it needs large amount of data to train, Varity of speakers. Continuous speech is more difficult to handle because of a variety of effects. First, it is difficult to find the start and end points of words. The production of each phoneme is affected by the production of surrounding phonemes, and similarly the start and end of words are affected by the preceding and following words. The recognition of continuous speech is also affected by the rate of speech (fast speech tends to be harder) (Abraham Woubie Zewoudie,2011, Ankita Chugh, Poonam Rana, Suraj Rana, 2014).

Connected is a halfway point between isolated word and continuous speech recognition. Allows users to speak multiple words.

2.5 Use and Application areas of Speech Recognition

As we mentioned before Speech recognition system the processes of converting spoken words, phrases or sentences into a text and also perform commands. In this case we can apply Speech Recognition in many areas of Daley life activities like: in industries such as business, banking,

marketing, and healthcare is quickly becoming apparent (Abd-ErrahimMaazouzi, 2019).

People with disabilities

People with disabilities can benefit from speech recognition programs. For individuals that are Deaf or Hard of Hearing, speech recognition software is used to automatically generate a closed-captioning of conversations such as discussions in conference rooms, classroom lectures, and/or religious services. Speech recognition is also very useful for people who have difficulty using their hands, ranging from mild repetitive stress injuries to involve disabilities that preclude using conventional computer input devices. Speech recognition is used in deaf telephony, such as voicemail to text, relay services, and captioned telephone. Individuals with learning disabilities who have problems with thought-to-paper communication (essentially they think of an idea but it is processed incorrectly causing it to end up differently on paper) can possibly benefit from the software but the technology is not bug proof. Also the whole idea of speak to text can be hard for intellectually disabled person's due to the fact that it is rare that anyone tries to learn the technology to teach the person with the disability. This type of technology can help those with dyslexia but other disabilities are still in question. The effectiveness of the product is the problem that is hindering it being effective (Pavel Kr´al, 2018).

Health Care: - Medical documentation

In the health care sector, speech recognition can be implemented in front-end or back-end of the medical documentation process. Front-end speech recognition is where the provider dictates into a speech-recognition engine, the recognized words are displayed as they are spoken, and the dictator is responsible for editing and signing off on the document. Back-end or deferred speech recognition is where the provider dictates into a digital dictation system, the voice is routed through a speech-recognition machine and the recognized draft document is routed along with the original voice file to the editor, where the draft is edited and report finalized. Deferred speech recognition is widely used in the industry currently (Jungsuk Kim and Wonyong Sung, 2012).

Therapeutic use

Prolonged use of speech recognition software in conjunction with word processors has shown benefits to short-term-memory re-strengthening in brain AVM patients who have been treated with resection. Further research needs to be conducted to determine cognitive benefits for individuals whose AVMs have been treated using radiologic techniques (Pavel Krbec, 2018, Jungsuk Kim and Wonyong Sung, 2012).

Military: - High Performance Fighter aircraft

Substantial efforts have been devoted in the last decade to the test and evaluation of speech recognition in Fighter aircraft. Of particular note have been the US program in speech recognition for the Advanced Fighter Technology Integration (AFTI)/F-16 aircraft (F-16 VISTA), the program in France for Mirage aircraft, and other programs in the UK dealing with a variety of aircraft platforms. In these programs, speech recognizers have been operated successfully in Fighter aircraft, with applications including: setting radio frequencies, commanding an autopilot system, setting steer-point coordinates and weapons release parameters, and controlling flight display. It was evident that spontaneous speech caused problems for the recognizer, as might have been expected. A restricted vocabulary, and above all, a proper syntax, could thus be expected to improve recognition accuracy substantially. Speaker-independent systems are also being developed and are under test for the F35 Lightning II (JSF) and the Alenia Aermacchi M-346 Master lead-in Fighter trainer. These systems have produced word accuracy scores in excess of 98% (Pavel Krbec, 2018, Jungsuk Kim and Wonyong Sung, 2012).

Training air traffic controllers

Training for air traffic controllers (ATC) represents an excellent application for speech recognition systems. Many ATC training systems currently require a person to act as a "pseudo-pilot", engaging in a voice dialog with the trainee controller, which simulates the dialog that the controller would have to conduct with pilots in a real ATC situation. Speech recognition and synthesis techniques offer the potential to eliminate the need for a person to act as pseudo-pilot, thus reducing training and support personnel. In theory, Air controller tasks are also characterized by highly structured speech as the primary output of the controller, hence reducing the difficulty of the speech recognition task should be possible. In practice, this is rarely the case. The FAA document 7110.65 details the phrases that should be used by air traffic controllers. While this document gives less than 150 examples of such phrases, the number of phrases supported by one of the simulation vendor’s speech recognition systems is in excess of 500,000. The USAF, USMC, US Army, US Navy, and FAA as well as a number of international ATC training organizations such as the Royal Australian Air Force and Civil Aviation Authorities in Italy, Brazil, and Canada are currently using ATC simulators with speech recognition from a number of different vendors (Pavel Krbec, 2018,).

Telephony and other domains

ASR is now commonplace in the field of telephony and is becoming more widespread in the field of computer gaming and simulation. In telephony systems, ASR is now being predominantly used in contact centers by integrating it with IVR systems. Despite the high level of integration with word processing in general personal computing, in the field of document production, ASR has not seen the expected increases in use. The improvement of mobile processor speeds has made speech recognition practical in smartphones. Speech is used mostly as a part of a user interface, for creating predefined or custom speech commands (TokunboOgunfunmi and Koji Seto, 2016, Ankita Chugh, Poonam Rana, Suraj Rana,2014).

2.6 Review of works done on Amharic Automatic Speech Recognition

Kinfe (2002) developed sub-word based isolated word recognition systems for Amharic Using HTK. The sub-word units used in the experiment are phones, triphones and CV-syllables. It considered 20 phones (out of 39) and 104 CV syllables, which are formed using the selected phones. Speech data of

170 words, which are composed of the selected sub-word units, have been recorded by 20 speakers (speech of 15 speakers for training and the remaining for testing). Speaker dependent phone-based and triphone-based systems have an average recognition accuracy of 83.07% and 78% respectively. Phone-based and triphone-based speaker independent systems have an average recognition accuracy of 72% and 68.4% respectively. In addition, a comparison of the different sub-word units revealed that the use of CV syllables has led to relatively poor performance.

Zegaye (2003) investigated the possibility of developing large vocabulary, speaker independent and continuous speech recognizer for Amharic language. The recognizer was developed using both phone-based and tri-phone based recognizers using HTK. For the training and testing of its recognizers, it used the speech data of 8000 sentences read by 80 people. In order to support the acoustic models, a bigram language model was constructed. In addition, pronunciation dictionary was prepared and used as an input. Since the recognizer was meant to recognize large vocabulary and continuous speech, phonemes were opted as the basic unit of recognition. The best recognizer was a tri-phone based recognizer which has 76.2%-word recognition accuracy.

Martha (2003) explored the possibility of developing an Amharic speech input interface to command and control Microsoft Word. It required a speech recognizer and a communication interface

between the recognizer and the application. 50 command words were selected from different menus (File, View, Insert, Tools, Table, Window, and Help), translated to Amharic and used to develop the prototype system. To develop and test the required Amharic speech recognition system, speech data were recorded from 26 people (10 females and 16 males) in the age range of 20 to 35. 76.9% of the recorded data were used to train the recognizers and the remaining data were used for testing the performance of recognizers. To test the performance of the system, 18 randomly selected command words were given to 6 people (3 command words for each) and these people were asked to command Microsoft Word orally. The system performed 16 commands accurately and only two command words were wrongly recognized and thus Microsoft Word performed wrong actions.

2.7 Related Work

These referred papers are used as an input, not as a comparison to our work. most of them are done for other languages, and also they are very relevant to the topic we are working on.

A. *Ranjodh Singh, Hemant Yadav, Mohit Sharma , Sandeep Gosain1, Rajiv Ratn Shah* (2019 IEEE) **"AUTOMATIC SPEECH RECOGNITION FOR REAL TIME SYSTEMS**".

Automatic Speech Recognition (ASR) systems have proven to be a useful tool to perform various day to day operations when used along with systems like Personal AI Assistants. Various industries require ASR to be trained in their domains. Music on Demand (MoD), over IVR, is one such industry where the user interacts with the dialogue system to play music using voice commands only. Domain adaptation of the model is expected to perform well on this domain as systems trained on public datasets are very generic in nature and not very domain specific. To train the ASR for MoD, experiment with the HMM-based classical approach and DeepSpeech2 on Voxforge dataset has been done to train the ASR system.

Both the models were first trained on publicly available Voxforge dataset to learn generic acoustic features. Voxforge dataset is a collection of transcribed speech files via user uploads. The dataset is unbalanced and has a lot of male examples. Furthermore, it has a variety of dialects of English language, from Indian English to Australian English. The dataset is available at two sample rates: 8k and 16k, the researcher uses the 8k version. The testing set for both the classical and Deep Models is a part of this Voxforge dataset and is exactly the same for both models.

Acoustic features are being used by various applications like ADVISOR combined with other features like geographic and visual to recommend personalized video soundtrack recommendation. Audio and text features are also being used in cross-modal correlation learning. Once the system trained DeepSpeech2 to learn generic acoustic features, then fine-tune the model on the data to adjust the acoustic and language model as per the domain.

Classical models like HMM-GMM models, Monophone and Triphone Training, MMI discriminative training are extremely intuitive in the way they make predictions and provide a faster inference time with explainable theory, but depend on the underlying assumptions about the data and require a lot of hand crafting. Future work to improve upon the brittleness of classical models can be done by leveraging the approximation power of Deep Neural Networks to estimate the probabilities in place of GMMs and using the alignments of the best HMM-GMM models as inputs to the various DNN-HMM models. Convolution Neural Networks and Long Short Term Memory networks are extremely successful at modelling temporal dependencies and hence using those in conjunction with Classical methods should result in better performance. End to end deep learning methods presents the exciting opportunity to improve the ASR systems as the data and computation power increases. The results also prove that compared to classical approaches the WER is substantially less for Deep learning methods. Also, the noise is better handled by the convolution layers. Without a good LM we cannot achieve a WER close to 10.

B. *Maazouzi, 2A. Laaroussi, 1N. Aqili, 1M. Raji and 1A. HammouchLRGE, ENSET, Mohammed V University in Rabat, Morocco ENSEM, Hassan II University in Casablanca, Morocco,* (2016) IEEE

"Speech Recognition System using Burg Method and Discrete Wavelet Transform"

Burg Method and Discrete Wavelet Transform are used to Presents an automatic recognition system of the English spoken digits using a parametric method of power spectrum density estimation. The signal is pre-processed by using the endpoint detection algorithm and then Discrete Wavelet Transform is applied to extract the approximation coefficients. To estimate the power spectrum density, Burg method is used. Daubechies wavelet family have chosen for features extraction step, because it gave a good result compared with the others families. The accuracy obtained by the developed scheme, the time consumption is significantly optimized i.e., the vector length characterizing each digit model was taken into consideration.

C. *KIRAN.R, NIVEDHA.K, SUBHA.T and PAVITHRA DEVI.S Department of information*

Technology,Sri Sai Ram engineering collage,Chennai-44, Tamil Nadu (2017) IEEE

"Voice and speech recognition in Tamil language".

The intention of this paper is to create a voice and speech recognition system in smart phones that recognizes voice and captures the speech data in Tamil and stores and converts the captured speech as text in Tamil language itself. This can be used in voice dialing, sending SMS by saying out the message and the captured message is sent to the recipient in Tamil. This paper describes also achievements of Automatic Speech Recognition System in many applications. Among them, Template Matching techniques like Dynamic Time Warping (DTW), Statistical Pattern Matching techniques such as Hidden Markov Model (HMM) and Gaussian Mixture Models (GMM), Machine Learning techniques such as Neural Networks (NN), Support Vector Machine (SVM), and Decision Trees (DT) are most popular. For the proposed system, HMM technique is used. HMM is a statistical pattern matching approach for speech recognition. It is the dominant technique for speech recognition based on statistical acoustic and language model is HMM. It uses the automatic learning procedures to model the variations of speech. A HMM for a sequence of words or phonemes is made by concatenating the individual trained HMM for the separate words and phonemes. It works by starting at upper left corner of trellis and generate observations according to permissible transitions and output probabilities. The output and transition probabilities define a HMM. The HMM addresses three problems to perform ASR and the solutions are given by the following three algorithms.

i. Computing the overall likelihood generating strings of observations from HMM by using the *Forward algorithm.*

ii. Decoding the most likely state sequence/best path from HMM by using the *Viterbi algorithm.*

iii. Learning parameters (output and transition probabilities) of HMM from data by using Baum Welch also called *Forward-Backward algorithm*.

D. *Basem H. A. Ahmed and Ayman S.Ghabayen, Department of computer science and Technology University collage of Science and Technology Gaza, Palestine*

"Arabic Automatic Speech Recognition Enhancement" (2017).

This paper proposes three approaches for Arabic automatic speech recognition. For pronunciation modeling, it proposes a pronunciation variant generation with decision tree. For acoustic modeling, it proposes the Hybrid approach to adapt the native acoustic model using another native acoustic model. In this paper, the acoustic frontend considered as a decode stage which involves the signal processing

to digitize analog signal and convert it to a sequence of feature vectors providing a compact representation of the given input signal. The Acoustic model (AM) represents acoustic features for phonetic units to be recognized. The proposed method for improving the accuracy of Arabic ASR system uses some methodologies. Among them, Pronunciation modeling, Pronunciation adaptation, Pronunciation Variant Generation with Decision Tree, Acoustic modeling and Language Modeling. The result shows that the proposed approach outperforms the baseline by relative reduction in WER. In this work Arabic language resources have been used to improve Arabic ASR.

E. *Kim, Euisung, Minnesota State University Mankato*

"A Wavelet Transform Module for a Speech Recognition Virtual Machine" (2016).

explores the trade-offs between time and frequency information during the feature extraction process of an automatic speech recognition (ASR) system using wavelet transform (WT) features instead of Mel-frequency Cepstral coefficients (MFCCs) and the benefits of combining the WTs and the MFCCs as inputs to an ASR system. A virtual machine from the Speech Recognition Virtual Kitchen resource (www.speechkitchen.org) is used as the context for implementing a wavelet signal processing module in a speech recognition system. It explored two different feature extraction methods, made comparisons, and looked for benefits of combining the two methods. Contributions include a comparison of MFCCs and WT features on small and large vocabulary tasks, application of combined MFCC and WT features on a noisy environment task, and the implementation of an expanded signal processing module in an existing recognition system.

F. *Michael MeleseWoldeyohannis, Laurent Besacier, Million Meshesha (1) Addis Ababa University, Addis Ababa, Ethiopia(2) LIG Laboratory, UJF, BP53, 38041 Grenoble Cedex 9, France* **"Amharic Speech Recognition for Speech Translation" (2016).**

This paper an attempt is made to experiment on Amharic speech recognition for Amharic-English speech translation in tourism domain. Since there is no Amharic speech corpus, the researchers developed a read-speech corpus of 7.43hr in tourism domain. In the study, The Amharic speech corpus has been recorded after translating standard Basic Traveler Expression Corpus (BTEC) under a normal working environment. In ASR experiments, phoneme and syllable units are used for acoustic models, while morpheme and word are used for language models. The ASR experiments have been conducted at different acoustic, lexical and language model units. In all experiments, Kaldi tool and SRILM are used for speech recognition and for language model respectively. Based on the experiment result, Encouraging ASR results are achieved using morpheme-based language models and phoneme-

based acoustic models with a recognition accuracy result of 89.1%, 80.9%, 80.6%, and 49.3% at character, morph, word and sentence level respectively.

G. *SevcanAytaçKorkmaz and FurkanEsmerayElectric and Energy Department Munzur University Tunceli, Turkey*

"Quality Lignite Coal Detection with Discrete Wavelet Transform, Discrete Fourier Transform, and ANN based on k-means Clustering Method" (2018) IEEE.

The lignite coal data in the KalburcayO area of the Sivas-Kangal Basin have been used. This original data obtained from KalburcayOarea of the Sivas-Kangal Basin consists of 66 observations in the lignite coal area, including lignite quality parameters such as moisture content, ash, sulfur content and calorific value. These lignite coal data's have been clustered in two group with k-means method according to calorie values. This clustering lignite coal data is classified by the Artificial Neural Network (ANN) classifier. In addition, Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) have been applied to coal data for ANN classifiers. DFT_ANN, DWT_ANN, and ANN classification success results are compared. The highest classification success rate was found by DWT_ANN method.

H. *Sevcan Aytac Korkmaz and Furkan Esmeray Electric and Energy Department Munzur University Tunceli, Turkey* **"Quality Lignite Coal Detection with Discrete Wavelet Transform, Discrete Fourier Transform, and ANN based on k-means Clustering Method" (2018)**.

The lignite coal data in the KalburcayO area of the Sivas-Kangal Basin have been used. This original data obtained from KalburcayO area of the Sivas-Kangal Basin consists of 66 observations in the lignite coal area, including lignite quality parameters such as moisture content, ash, sulfur content and calorific value. These lignite coal data's have been clustered in two group with k-means method according to calorie values. This clustering lignite coal data is classified by the Artificial Neural Network (ANN) classifier. In addition, Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT) have been applied to coal data for ANN classifiers. DFT_ANN, DWT_ANN, and ANN classification success results are compared. The highest classification success rate was found by DWT_ANN method.

The literature referred for this thesis work are summarized as follows. Most of the papers are used as an input. Amharic Speech Recognition related papers are used as a reference for the current state of Amharic Speech Recognition System and to identify the research gap between current speech recognition technology and the research conducted in this thesis work.

Most of Amharic Speech recognitions are isolated word recognition and Speaker defendant. In this case may the accuracy will be high. Although important for their intended purpose, but no accessible for every user. To make the service accessible, there must be research for speaker independent speech recognition system. The main resource for every SRs is speech data. In this thesis work 20 hr. audio is used to conduct speaker independent speech recognition and achieved minimum word error rate. To decrease the word error rate to minimum, more speech data is needed.

CHAPTER THREE

METHODOLOGY

This chapter presents the research procedure, solution methods, development, and algorithm being proposed to solve the complex problem of speech recognition on Amharic Speech Recognition. The following flow chart shows the general flow of processes.

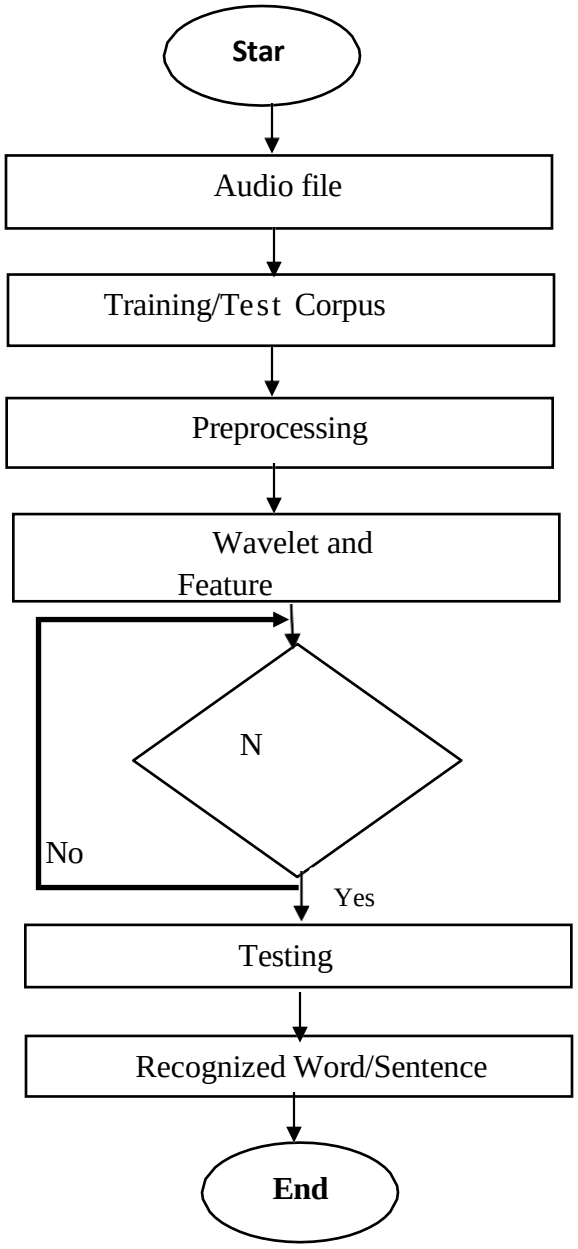

Figure 3.1 Flow chart of the proposed system

3.1 Collecting data samples

The data is collected from Amharic speakers with different accent and pronunciation. Speakers are required to speak (record) each Amharic sentence or word clearly, in low noise in background. The dataset was collected from ALFFA_PUBLIC (African Languages in the Field, Speech Fundamentals and Automation (http://alffa.imag.fr/)) which is working on Kaldi Speech Recognition Tool, by recording different speakers, recording from radios, and downloading videos from YouTube. After downloading and recording; it was necessary to make format change and break apart the file in the size required by the systems.

Corpus preparation

One of the most fundamental resources for any ASR system and development is speech and text corpora. That is the main reason for the preparation of an Amharic Speech Corpus as a first step in exploring the possibilities of an Amharic Automatic Speech Recognition System. Having the text corpus, the next important step in corpus preparation is recording the speech. For continues speech, a test corpus has to be developed that can be read and recorded. It is the collection of audio files and the transcribed text file of the audio used for the ASR system. In this work, a total of 9340 sentences with a length ranging from 2 to 18-word length have been recorded, downloaded and processed. Sample transcribed sentences for training are shown below in figure 3.2

Figure 3.2 Training Corpus Sample

The transcribed sentence is assigned to the corresponding audio file name without file extension as shown in the figure 3.3 below.

	A	B
1	ያንደኛ ደረጃ ትምህርታቸው ን ጉንደር ተ ም ረቀል	tr_1_tr01001
2	የተ ለቀቁት ምርኮኞች በ አዞበቢያቸው ሰላማዊ ኑር እንዲ ናሩ የ ትራንስፓርት ና መጓጓዣ ገንዘብ ተሰጥቷቷ ቸው መሸኝታቸው ን አመልክቱ በ የ ዞ ናቸው እንደ ደረሱ መቃቋሚያ እንደሚ ሰጣቸው ም አስ ታውቋል	tr_2_tr01002
3	በ አዲስ አበባው ስታዲየም በ ተሰሩዱ ት ሁለት ግጥሚያ ቆች በ መጀመሪያ የ ተጉኙ ት መይን ና ሙ ዝር ሲሚጎች ሲ ሆኑ በ ውጤቱ ም ዮስት ለ ዮስት ተለያይ ተዋል	tr_3_tr01003
4	ወሬው ን ወሬ ያደረጉ ምስጢረ ኞች ናቸው	tr_4_tr01004
5	ኢትዮጵያዊ ቷ በ ብሄራዊ ባህላዊ አለባበስ ከ አለም እጎደኝነት ን ተቀላጀ ች	tr_5_tr01005
6	ከ ትምክህት እንሳይ ቆጠር ብን እንጂ በ አለም ታሪክ ውስጥ በ ነፍች ያል ተረገጠ ች እገር ኢትዮጵያ ና ት	tr_6_tr01006
7	አሀዞቹ የኢርትራ ዜጐች ና የ ሻአቢያ ደጋፊዎች ናቸው	tr_7_tr01007
8	አናንተ ም መቀበሪያ እንዳ ታጡ ተጠንቀቁ	tr_8_tr01008
9	እንቶኒሊ በ አጸ ምንሊክ ፊት የ ፊጸመው ይፍረት በ ኢጣሊያ ን ምክር ቤት አስተ ቸው	tr_9_tr01009
10	ግን ወደ ኃላው ሳይ ኢሳያስ እንደ ልማ ሳቸው ሁሉን ም የ መልከ ፍ ዲፕሎማሲ ያቸው አስራሊል ንም ያስ ወር ፋቸው ጀመር	tr_10_tr01010
11	ከ የ አቅጣጫ ው አየ ደረጉቸው ያሉ መሪጀዎች አሳሳቢ ችግሮች አየ ደረሱ መሆናቸው ን የሚ ጠቁሙ መሆናቸው ን ፕሬዝዳንቱ ተናገሪ ቸል	tr_11_tr01011
12	ከ ማወቁ በፊት እንደ ተጨበበ ገበያ እንሳይ በታተን ያህ ነው አጀንሳ ችን ሊላ አጀንሳ የ ለ ንም	tr_12_tr01012
13	ኢትዮጵያ ም ሰራዊቷ በ ኢርትራ አንደሚ ገኝ አል ካህ ችም	tr_13_tr01013
14	ላቁእት ትምህርት ቤት መንገድ ና ሆስፒታል ተ ገንቢቷል	tr_14_tr01014
15	አስሬያን በ ወገፍ አንድ ለ ዘር አሸንፎ ሶስት ነጥብ ያ ሃል	tr_15_tr01015
16	ሁሴን እይዲድ እንደ ገለጹት ኢትዮጵያ ሁኔታዎች ከ ተመቾቱ ላት ሶማሊያ ን ከ መውረር ስለማ ት መለስ ከ አርቆ ይልቅ ሶማሊያ ን ከ ኢትዮጵያ ና ኢትዮጵያ ካ ደራጀ ቻቸው ሀይሎች ለ መጠበቅ መዘጋጀት ያስፈልጋ ል	tr_16_tr01016
17	በ ሙስና ከስ የ ተመሰረተባቸው ኑንጀዎች መንናኛ ብዙሀን ን ከሰሱ	tr_17_tr01017
18	ኢትዮጵያ የመን ና ሱማን ኢሳያስ ን ለ ማውረይ ተቧማም ተዋል ተ ባለ	tr_18_tr01018
19	ለማ ን አቢት ማለት አንደሚ ቻል ም ግራ ግቡዶ ናል በ ማለት ኑንጀ ቻቸ ተሰጥ ተውን ገልጸዋል	tr_19_tr01019
20	በ ኢትዮጵውያ ን ብቻ ሳን ወሰን የ ኢትዮጵያ ወጎጀች የሆኑ የ ኢትዮጵያ ታሪክ እቆቋቻች ም በ ውጪ የሚገኙ በዚህ ስራ ላይ ተሳታፉ እንዲሆኑ ጥረት አደርገ ናል	tr_20_tr01020
21	ኢትዮጵያ ውስጥ የ ህገት ኢርትራውያን ም ተመሳሳይ አይል ሃሪቶቸው	tr_21_tr01021
22	የ ደህንት ኢሚግሬሽን ና ስደተኞች ባለስልጣን ሰራተኞች በ በኩላቸው ተዓጉቦ ሰኑሉ ን እንዲ ያሟሉ ለ ቀዩ መስቀል ኮሚቴው ማስታወ ቻቸው ን ተናግረ ቸል	tr_22_tr01022
23	ሁለት መቶ ሺ ያህሉ ደግሞ አብዛኛዎቹ ሲ ሰደዱ ቀሪዎቹ ሞ ተዋል ወይን ም ተ ሸሽገ ዋል	tr_23_tr01023
24	ዛሬ ም ደግ መ ን ያህኑ እን ላለን	tr_24_tr01024
25	በ ቦሌ አለም አቀፍ አውሮፕላን ማረፊያ ጸረ ናርኮቲክ የ ቁጥጥር ክፍል አደገኛ አዶች አዘዋዋሪ ዎች ኢትዮጵያ ን መሸጋገሪያ አንዳያ ደርጓት ማሳሰቡ ን በ ዘኽው እ ስፍሯል	tr_25_tr01025
26	ዘጠና ሁለት ኢትዮጵያውያ ን የ ደረሰብት ጠፋ	tr_26_tr01026
27	መረጃ ያልተገኘ ባቸው በታቸው ናቸው በ ማለት ምላሽ ሰጥተዋል	tr_27_tr01027
28	ኮንፍረንሱ ን የ መሩት ስር ቺሊ ያም ራሪ አሚስ ፎሪ ም ለአፍሪካ ኢንቨስተመንት አዲስ ራእይ አንደሆነ ገልጸዋል	tr_28_tr01028
29	መላለክት ያጀቡ ት ህጻን ቦሪና ህጻናት ን የሚጠበቁ ና የሚንዘበበ ዊቸው መላለክት አሉ ይባላል	tr_29_tr01029
30	አውቅና ን መ ንጽፍ መጥፎ ጉን አንዳለው በሚገባ ተረድ ቻቸዉ	tr_30_tr01030
31	አውቅና ን ማግኘቱ ለ አኔ ትልቅ ኩብር ነው	tr_31_tr01031
32	ም ን ለማ ለት ነው ገልጽ አድርገው	tr_32_tr01032
33	በዚያ በ ተጨማሪ የ ስልጠና ውን ሂደት የሚ ያሻሽል ላቸው ያ ሻሉ	tr_33_tr01033
34	ጨዋታው ቁጥሉ ፔናልቲ ውስጥ ቺላቪርት ኒስ ያገኝ ና ሳ ያውቅ ለ ፓሊርሞ ያቀ ብለዋል	tr_34_tr01034
35	አንዋር ጊዮርጊስ ሊ ዝነ ነው	tr_35_tr01035
36	ሆን ተብሎ ም አየተደረገ ብኝ ነው	tr_36_tr01036
37	ጨዋታው የሚ ከናነው ሚያዝያ አስራ ስምንት ቀን ነው	tr_37_tr01037
38	አንጄር ግዛው ተክለ ማሪያም ገልጽ ባልሆነ ምክንያት ከ ተጫዋቾቹ ጋር ሳይ ተዋወቁ ቀር ተዋል	tr_38_tr01038
39	ደንቡ ን ያጸደቁ ው የ ፌደራል አ ስ/ርት ምክር ቤት ነው	tr_39_tr01039
40	ያን ን ማ ፕርግራም ያወጡ ት አ ዮ አነ ዘርር አቶ ያይነታቸው ተሰማ ናቸው	tr_40_tr01040

Figure 3.3 Training transcription with audio file name

As of training processes the test or validation file also prepared as Validation Corpus and Validation Transcription as shown figure 3.4 and 3.5 respectively below.


```
'01_d501021.wav", "text": "የ ኢንተርኔት አገልግሎት ንም በ ተመለከተ በ ከልሎች በ ዘኖች ና በ አዲስ አበባ በ ተለያዩ ቦታ ቃች የ አገልግሎት ማእከሎ ችን ለ ማቋቋም መታቀዱ ን አብራር ተዋል", "duration": 12.544}
'01_d501022.wav", "text": "በዚህ ም አነስተኛ መነኖ ም ችን በ ማስፋፋት አገራቱ ን በ ተደጋጋሚ ከሚ ያጠቃ ት ድርት ና የ ምግብ እህል አጥረት ለ ማቃለል እንደሚ ቻል ጠቁ መዋል", "duration": 12.544}
'01_d501023.wav", "text": "ይህ እንሳይሆን በ ህይወታቸው መበሳእትነት ያስገኙ ትን የ ኢትዮጵያዊ ነት ኩበር አንግ በ ን ለ ቅድስት ሀገራችን አንድነት ና ሉአላዊነት ሰላም ና ብልፅግና በ ሁበረት እንቁ ም", "duration": 13.696}
'01_d501024.wav", "text": "የመጀመሪያ ው ነጥብ በ ተቃዋሚ ሀይሎች ሰፊር በ ጋራ አቋም ላይ በ ጋራ ለ ማቆም ያየነው ፅናት አነስተኛ መሆኑ ነው", "duration": 10.752}
'01_d501025.wav", "text": "አቅም በሚ ፈቅደው መሰረት አስቀድሞ መዘጋጀት ስለሚ ፈልጉ የሚ ያመራጥር እይሆን ም ስትል የ አሚረባ ድምፅ ዘና ዘጋቢ ገልፃ ለች", "duration": 11.392}
'01_d501026.wav", "text": "ከ ለ ቀስተኞቹ መከከል አብዛኞቻቸ ዘበ ዛሬ ተገለጠ ሀይለስላሴ ማፍን አለማችን አብቃችን የ አፍሪካ አሀት የ አለም አብት በ ማለት ነበር ዘዘና ቸውን የሚገልፁ", "duration": 16.0}
'01_d501027.wav", "text": "በዚህ ጋር ለ ሱብአሚ አገልግሎት መሰለ ፉ ም ውጤታማ መሆኗ ም አድናቆት ን አትርፎ ላታል", "duration": 8.064}
'01_d501028.wav", "text": "የ ኮሚሽኑ ውሳኔ ና የቤተ ክህነቱ ቋሚ ሲኖዶ ም", "duration": 4.736}
'01_d501029.wav", "text": "ይህንን ም ባድመን ና ሽራሮ ን በ መውረር እ ውን አ ርጓል", "duration": 5.376}
'01_d501030.wav", "text": "ጋዜጠኞች ን መለያየታቸው ብዙም ች ን አሳዝኗ ል", "duration": 4.864}
'01_d501031.wav", "text": "ባለፈው ሰኞ አ ልት ደግሞ በ ቴሌቭቪዥን ሊላ ሽልማት ሲ ሸለም ተ መልከት ቡ", "duration": 7.168}
'01_d501032.wav", "text": "አጠቃላይ ወጪው አስራ ስምንት ሺ ዶለር እንደሆነ ለማወቅ ተችሏል", "duration": 5.504}
'01_d501033.wav", "text": "ሌሎቹ በ ሙሉ ጠነ ኞች ናቸው", "duration": 2.944}
'01_d501034.wav", "text": "መለስ ና ኢሳያስ እርስ በ እርስ ለ መገለባበጥ እየ ሞከሩ ነው", "duration": 5.12}
'01_d501035.wav", "text": "የ አየር ሀይላችን አውሮፕላኖች ም ግፋኝ ቸውን በ ተገቢው ፈፅመው በ ሰላም ተ መልሰዋል ሲል የመንግስት ቃል አቀባይ ፅፊት ቢት መግለጫ አስ ታውቋል", "duration": 12.672}
'01_d501036.wav", "text": "ከ ሚ ኞቹ ጋር የነበሩ ሁለት ታጣቂዎች በ ሰላም አጀቸው ን መከጠታቸው ንም የ ፖሊስ መግለጫው አስሎ ገልፆ ል", "duration": 8.192}
'01_d501037.wav", "text": "ችግሮቹ ን ደግሞ ማስወገድ ነበረ ባቸው", "duration": 4.224}
'01_d501038.wav", "text": "አጋፋሪ ቃቻቸው ባ አዳ ኑም አቋማቸው ለ ነርሱ ወር ቶ ና ሙ ቡቶ በ ሰጡት ድጋፍ አይ ለይም", "duration": 9.6}
'02_d502021.wav", "text": "ይሄኔ መለስ ነቃ አለ", "duration": 2.56}
'02_d502022.wav", "text": "በዚህ ከልል ለ በርካታ ዜጐች ም የበራ አይል ይ ፊጥራል የሚሉ ሲ ኖሩ በተለይ ለ ሶማሊ ያውያን አይሉ ሰፉ እንደሆነ ተጠቅ ሷል", "duration": 9.344}
'02_d502023.wav", "text": "ድፍረት ባይሆን ብኝ በህተት ነው", "duration": 2.816}
'02_d502024.wav", "text": "በ ሀገራችን የ ቅርብ ዘመን ታሪክ ውስጥ አራት አዲስ ምእራፍ ከፋች ው ነቆች መጥቀስ ይቻላል", "duration": 6.784}
'02_d502025.wav", "text": "የ ቢራ ው ኢንዱስትሪ ግን የ ታክስ ችሳባ ቢደረግ ለትም ም ንም አይነት የ ዋጋ ለውጥ ል ደንባቸዉ አላደረገ ም", "duration": 7.552}
'02_d502026.wav", "text": "ህወሀት ኢህአዴግ ታድ ሻለሁ ከ ማለቱ ና ቀውሱ በ መፈጠሩ በፊት ፖሊስ በ ፍትህ ሚኒስቴር ስር እንደ ነበር ሲ ታወት ፍትህ ሚኒስትሩ ም የ ህወሀት ሰው እንዳል ነበሩ ይታ ው ሳል", "duration": 10.752}
'02_d502027.wav", "text": "አሁን ለምን ድምፃቸው ጠፋ ሁለቱ ኣገሮች በ ጀመሩት ጦርነት የበሉ የ ሰጋ ው የ አሚረባ መንግስት ነው", "duration": 6.912}
'02_d502028.wav", "text": "ሁሴን አይዲይ በ አሊትሀይ ላይ አቋም ወሰዓ ስምምነት ም አደረጉ", "duration": 4.992}
'02_d502029.wav", "text": "በዚህ እንደ ተሰማሙ ለማ ን ት ሰጥ ለማ ን ሲ ባል ለዚህ ማ ካሳ አለ እይደል እንደ የ ኮ ሶ ሻሎ ልጅ የሚ ል ሀሳብ ተሰማ", "duration": 9.472}
'02_d502030.wav", "text": "ሮናልዶ መናገር የ ፈለገው ግን ህመሙ ሙሉ ለ ሙሉ ለ መዳን ጊዜ የሚ ፈልግ ግን የሚ ድን ለማ ለት ነበር", "duration": 7.296}
'02_d502031.wav", "text": "በ አስተያየቱ ትስማማ ለህ ተብሎ ሲ ጠየቅ አስልጣኝ ሀጎብ ይህ አይነት አጋማሚ በ መላው አለም ላይ ሲ ከስት ተስተ ው ሏል በ ማለት መልሷ ል", "duration": 9.216}
'02_d502032.wav", "text": "አነልካ ከ ሺ ማይከል ጋር በ ተደጋጋሚ ተገናኝ ቶ ያየ ከ ናቸው ኒሰች የሚያሰ ቆጩ ናቸው", "duration": 6.912}
'02_d502033.wav", "text": "በ ጀርመን የሚገኙ ወጣቶች ም ለ ብሄራዊ ቡድን እድል ስለማይ ሰጣቸው ዘግኔታቸው ን እየ ለወጡ ይገኛሉ", "duration": 7.936}
'02_d502034.wav", "text": "አቶ አለማየሁ ዘገየ በ ዕረ ሙናና ኮሚሽን ከበ ላይ መከተታቸው ን ከ ጋዜጦች ላይ ከማን በ ባቸው ውጭ የ ፍርድ ቤት መጥሪያ እንሳል ደረሳቸው መናገራቸው ም ተጠቁ ሟል", "duration": 10.368}
'02_d502035.wav", "text": "ከ ኢትዮጵያ ጋር ብን ቀላቀል ብዙ የሚ ለወጥ ነገር የ ለ ም", "duration": 4.096}
'02_d502036.wav", "text": "የ ልን ግሎ ኢትዮጵያ ን ማህበር ኢርትራ ሰራዊቷ ን እንድ ታስወጣ ጠየቀ", "duration": 5.376}
'02_d502037.wav", "text": "እኔ ራሴ የ ት ባታ ሆኜ በራዬ ን እንደም ሰራ አላውቅ ም ማለታቸው ን ከ ቢቢሲ በሚ ተላለፈው ፎከስ ኦን አፍሪካ ፕሮግራም ላይ አስደም ጧል", "duration": 9.856}
'02_d502038.wav", "text": "የ ህገር አስተዳደር ሚኒስትሩ ልደ ረዳት በ መሆን ካ ገለፁ በኋላ በ ቴሀራን ና በ ቴልአቪቭ ጀርመን ኤምባሲዎች ውስጥ ተመድበው አገልግ ለዋል", "duration": 9.728}
'03_d503021.wav", "text": "በ አቶ መለስ የተጠራ ው ይህ የ ጦር ጀኔራሎች ስብሰባ እሁን ም እንደ ቀጠለ ነው ተ ብሏል", "duration": 4.864}
'03_d503022.wav", "text": "ባሊዶግሊ የ ኢትዮጵያ ጦር ተቆጣጥሮ ታል የ ተባለው የ አይሮፕላን ማረፊያ የሚ ገኝበት ከተማ ነው", "duration": 5.632}
'03_d503023.wav", "text": "ሻለቢያ እ ከ ሽፍ ኒዛቸው የሚ ላቸው ሰምንቱ የ ደርግ የ ዘመቻ ዙ ሮች ሙ ና ሆነው የ ቀሩት በ ወያኔ ተሞጊዶች መስየሳትነት እንደሆነ ይነገራ ል", "duration": 7.808}
'03_d503024.wav", "text": "በ ኢርትራ በኩል በ ተመሰሳይ መልኩ የተወሰኑ አላማዎች መሃደፋቸው ን እሙን ነው", "duration": 4.736}
'03_d503025.wav", "text": "የ አልበሽር መንግስት ከዚህ በኋላ የኢርትራ ሉአላዊነት የሚ ጎረር ና የ ባለ ስልጣናቱ ን ስብእና ከሚ ነካ ተግባር መቆጠብ እንደሚ ኖርበት ገልፀዋል", "duration": 8.832}
'03_d503026.wav", "text": "ይህ ሰው ውሀ ልማት አካባቢ በ ሚገኘው መጥሪያ ቤቱ የ ታሸገ በት ሲሆን እርሱ ም የ ደረሰበት አይ ታወቅ", "duration": 6.144}
'03_d503027.wav", "text": "እኛ ም ወስነ ት ይሆናል እያ ልን በኋላ በኋላ ነው ይህ የ ወጣለት", "duration": 4.736}
'03_d503028.wav", "text": "ከ ቁልቢ ወደ ሀረር ተ ወስደው መታሰሪቸው ከ መታወቁ በፅተቀር ይህ ዜና እስከ ተጠናከረ በት ጊዜ ድረስ ም ንም አይነት ለውጥ አለመኖሩ ን ውስጥ አዋቂ ምቹ ገልፀዋል", "duration": 7.808}
```


Figure 3.4 Validation Corpus Sample

	A	B
1	የ ኢንተርኔት አገልግሎት ንም በ ተመለከተ በ ክልሎች በ ዞኖች ና በ አዲስ አበባ በ ተለያዩ ቦታዎች የ አገልግሎት ማእከሎ ችን ለ ማቋቋም መታቀዱ ን አብራር ተዋል	01_d501021
2	በዚህ ም እነስተኛ መስና ዎችን በ ማስፋፋት አገሪቱ ን በ ተደጋጋሚ ከማ ያጠታ ት ድርቅ ና የ ምግብ እህል እጥረት ለ ማቃለል አንደሚ ቻል ጠቁ መዋል	01_d501022
3	ይህ አንዳያሆን በ ሀይወታቸው መስዋእትነት ያስገኙ ትን የ ኢትዮጵያዊ ነት ኩበር አንግ በ ን ለ ቅድስት ሀገራችን አንደነት ና ሉአላዊነት ሰላም ና ብልጽግና በ ሁበረት አንቁ ም	01_d501023
4	የመጀመሪያ ው ነጥብ በ ተቃዋሚ ሀይሎች ሰፊር በ ጋራ አቋም ላይ በ ጋራ ለ ማቆም ያዩነው ጽናት አነስተኛ መሆኑ ነው	01_d501024
5	አቅም በሚ ፈቅደው መሰረት አስቀድሞ መዘጋጀት ስለሚ ፈልጉ የሚ ያጠራጥር አይሆን ም ስትል የ አሚሪካ ድምጽ ዜና ዘጋቢ ገልጻ ለች	01_d501025
6	ከ ለ ቀስተኞቹ መካከል አብዛኛዎቹ ጸሀዩ ዛፊ ተገለጠ ሀይለስላሴ ማሩን አለማችን አባታችን የ አፍሪካ አባት የ አለም አባት በ ማለት ሃበር ሀዘና ቸውን የሚገልጹ	01_d501026
7	ከዚህ ጋር ለ ሰብአዊ አገልግሎት መሰለ ፌ ም ውጤታማ መሆኗ ም አድናቆት ን አትርፎ ላታል	01_d501027
8	የ ኮሚኒኝ ውሳኔ ና የቢተ ከሀሉ ተታውሞ	01_d501028
9	ይህንን ም ባድመን ና ሽራሮ ን በ መውረር አ ውን አ ርዓል	01_d501029
10	ጋዜጠኞች ን መለያየታቸው ብዙዎች ን አሳዝኗ ል	01_d501030
11	ባለፈው ሰኞ እ ለት ደግሞ በ ቴሌቭዥን ሌላ ሽልማት ሲ ሸለም ተ መልከት ቡ	01_d501031
12	አጠቃላይ ወጪው አስራ ስምንት ቪ ዶላር አንደሆነ ለማወቅ ተችሏ ል	01_d501032
13	ሌሎቹ በ ሙሉ ጤነ ኞች ናቸው	01_d501033
14	መለስ ና ኢሳያስ አርስ በ አርስ ለ መገለባበጥ አየ ሞከሩ ነው	01_d501034
15	የ አየር ሀይላችን አውሮፕላኖች ም ግዳጃ ቸውን በ ተገቢው ፈጽመው በ ሰላም ተ መልሰዋል ሲል የመንግስት ቃል አቀባይ ጽፊት ቤት መግለጫ እስ ታውቋል	01_d501035
16	ከ ሚ ኞቹ ጋር የነበሩ ሁለት ታጣቂዎች በ ሰላም እጃቸው ን መስጠታቸው ንም የ ፖሊስ መግለጫው አከሎ ገልጿ ል	01_d501036
17	ችግሮቹ ን ደግሞ ማስወገጅ ነበረ ባቸው	01_d501037
18	አጋፋሪ ዎቻቸው ባ አዳ ኦም አቋማቸው ለ ነርሱ ሀር ቶ ና ሙ ቡቶ ከ ሰጡት ድጋፍ አያ ለያም	01_d501038
19	ያሄኔ መለስ ነቃ አለ	02_d502021
20	በዚህ ክልል ለ በርካታ ዘጌች ም የስራ አድል ይ ፈጥራል የሚሉ ሲ ናሩ በተለያ ለ ሰማሊ ያውያን አድሉ ሰፊ አንደሆነ ተጠቅ ሷል	02_d502022
21	ድፍረት ባይሆን ብኝ ስህተት ነው	02_d502023
22	በ ሀገራችን የ ቅርብ ዘመን ታሪክ ውስጥ አራት አዲስ ምአራፍ ከፋች ሁ ነቶች መጥቀስ ይቻላል	02_d502024
23	የ ቢራ ው ኢንዱስትሪ ግን የ ታክስ ቅነሳ ቢደረግ ለትም ም ንም አይነት የ ዋጋ ለውጥ ለ ደንበኞቹ አላደረገ ም	02_d502025
24	ህወሀት ኢህአደግ ታድ ጣለሁ ከ ማለቱ ና ቀውሱ በ መፈጠሩ በፊት ፖሊስ በ ፍትህ ሚኒስቴር ስር እንደ ነበር ሲ ታወቅ ፍትህ ሚኒስትሩ ም የ ህወሀት ሰው እንዳል ነበሩ ይታ ው ሳል	02_d502026
25	አሁን ለምን ይምጻቸው ጠፋ ሁለቱ አገሮች በ ጀመሩት ጦርነት የበለጠ የ ሰጋ ው የ አሚሪካ መንግስት ነው	02_d502027
26	ሁሴን አይዲድ በ አሊትሀድ ላይ አቋም ወሰዱ ስምምነት ም አደረጉ	02_d502028
27	በዚህ እንደ ተስማሙ ለማ ን ት ሰጥ ለማ ን ሲ ባል ለዚህ ማ ካሳ አለ አይደል እንደ የ ኮ ሶ ሻጭ ልጅ የሚ ል ሀሳብ ተሰማ	02_d502029
28	ርኅልዶ መናር የ ፈለገው ግን ህመሙ ሙሉ ለ ሙሉ ለ መዳን ጊዜ የሚ ፈልግ ግን የሚ ድን ለማ ለት ነበር	02_d502030
29	በ አስተያየቱ ትስማማ ለህ ተብሎ ሲ ጠየቅ አሰልጣኝ ሀጐስ ይህ አይነት አጋጣሚ በ መላው አለም ላይ ሲ ከሰት ተስተ ው ሷል በ ማለት መልሷ ል	02_d502031
30	እኔልካ ከ ቪ ማይክል ጋር በ ተደጋጋሚ ተገናኝ ቶ ያባ ከ ናቸው ኳሶች የሚያስ ቆጩ ናቸው	02_d502032
31	በ ጀርመን የሚገኙ ወጣቶች ም ለ ብሄራዊ ቡድን አድል ስለማይ ሰጣቸው ዘግኑታቸው ን አየ ለወጡ ይገኛሉ	02_d502033
32	አቶ አለማየሁ ዘገየ በ ጸረ ሙስና ኮሚሽን ክስ ላይ መከተታቸው ን ከ ጋዜጦች ላይ ከማን በ ባቸው ውጭ የ ፍርድ ቤት መጥሪያ አንዳል ደረሳቸው መናገራቸው ም ተጠቁ ሟል	02_d502034
33	ከ ኢትዮጵያ ጋር ብን ቀላቀል ብዙ የሚ ለወጥ ነገር የ ለ ም	02_d502035
34	የ እን ግሎ ኢትዮጵያ ን ማህበር ኤርትራ ሰራዊቷ ን አንድ ታስወጣ ጠየቀ	02_d502036
35	እኔ ራሴ የ ት ቦታ ሆኜ ስራዬ ን አንደም ሰራ አላውቅ ም ማለታቸው ን ከ ቢቢሲ በሚ ተላለፈው ፎከስ ኦን አፍሪካ ፕሮግራም ላይ አስደም ጧል	02_d502037
36	የ ሀገር አስተዳደር ሚኒስትሩ ልዩ ሪዳት በ መሆን ካ ገለገሉ በኋላ በ ቲሀራን ና በ ቲልአቪቭ የ ጀርመን ኤምባሲዎች ውስጥ ተመደባው እገልግ ለዋል	02_d502038
37	በ እቶ መለስ የተጠራ ው ይህ የ ጦር ጀኔራሎች ስብሰባ አሁን ም እንደ ቀጠለ ነው ተ ብሏ ል	03_d503021
38	ባሊ ዶግሌ የ ኢትዮጵያ ጦር ተቆጣጥሮ ታል የ ተባለው የ አይሮፕላን ማረፊያ የሚ ገኝበት ከተማ ነው	03_d503022
39	ሻአቢያ እ ከ ሸፍ ኳቸው የሚ ላቸው ስምንቱ የ ደርግ የ ዘመቻ ዙ ሮች መ ና ሆነው የ ቀሩት በ ወያኔ ተዋጊዎች መስዋእትነት እንደሆነ ይነገራ ል	03_d503023
40	በ ኤርትራ በኩል በ ተመሳሳይ መልኩ የተወሰኑ አላማዎች መነደፋቸው ን አመን ነው	03_d503024

Figure 3.5 Validation transcriptions with audio file name

3.2 Pre-processing

Speech is an analog signal “a continuous flow of sound waves and silence.” This form cannot be processed directly by a speech recognizer. Thus, there is a need for preprocessing the speech signal. The development of any speech recognition system consists the following steps for the construction of ASR system. The steps are, Pre-processing, feature extraction, decoding. In the development of an Automatic Speech Recognition system, pre-processing is considered as the first phase of other phases in speech recognition to differentiate the voiced or unvoiced signal and create feature vectors. Preprocessing adjusts or modifies the speech signal, x(n), so that it will be more acceptable for feature extraction analysis. The major factor to consider when it comes to speech signal processing is to check the speech, x(n) if is corrupted by some background or ambient noise, d(n), for example as additive disturbance.

$$x(n) = s(n) + d(n) \quad (3.1)$$

Where s(n) is the clean speech signal. In noise reduction, there are different methods that can be adopted to perform the task on a noisy speech signal. To apply the preprocessing process in this system, we apply the discrete wavelet transform to the speech signal and. We also make sure that the samples have equal length to prevent the training or learning process hindered by over fitting.

The testing samples also have equal length to help us evaluate the system as the NIST LRE series standards.

3.3 Feature Extraction

Speech feature extraction is the signal processing frontend which converts the speech waveform into some useful parametric representation. These parameters are then used for further analysis in various speech related applications such as speech recognition, speaker recognition, speech synthesis and speech coding. It plays an important role to separate speech patterns from one another. Because every speech and speaker has different individual characteristics embedded in their speech utterances. Features represent data and serve as input to the learner. Speech recognition methods derive features from audio, such as Spectrogram or Mel Frequency Cepstrum (MFCC), Discrete Wavelet Transform (DWT) and other feature extraction methods. A DWT transforms an audio signal into time domain and frequency domain; this means that it can analyze the signal with respect to time as well as frequency. In DWT, the audio signal is decomposed up to L level of approximate and detail coefficients, where L is any integer number.

DWT process is applied on the wavelet transform in the input speech signal some measures are obtained. The signal is decomposed into two frequency bands such as low frequency band (approximation coefficients) and high frequency band (detail coefficients). Wavelets are highly useful technique for speech recognition in terms of de-noising and feature extraction. Wavelet separates high frequency and low frequency sub bands. Low frequency consists of features or characteristics of signals and high frequency consists of noise content. Wavelet is to diagnose in to low frequency component in to maximum levels in this research considered as 8 levels. The low frequency band contains more useful information pertaining to the speech signal.

DWT is a left recursive binary tree structure (Https://En.Wikipedia.Org/Wiki/Deep_Learning, Y. Song, B. Jiang, Y. Bao, Si Wei and Li-R. Dai, 2013). The frequency band of required signal was chosen between 50 Hz to 8000 Hz. The decomposition was carried out until these required frequency components were obtained. It resulted in 32 frequency bands of which only 11 were chosen for our analysis. The selection of mother wavelet function has a prime role in the design of high quality speech coding. In this work, three different types of Daubechies wavelets (db4, db6 and db8) were analyzed. Features such as mean, standard deviation, power, maximum value, minimum value and a fusion of all the specified features were used to obtain the feature vector.

Audio is split into small blocks, such as 10 milliseconds, and each block is broken into its constituent frequencies. The advantage of applying the Mel-scale is that it approximates the nonlinear frequency resolution of the human ear (Kim-Y. E. Wong, 2014). As with any filter-bank based speech analysis technique, an array of band-pass filters are utilized to analyze the speech in different frequency bandwidths. In MFCC parameterization, the positions of the band-pass filters along the linear frequency scale are mapped to a Mel-scale according to:

$$Mel(f) = 2595log10(1 + \frac{f}{700}) \quad (3.2)$$

Triangular filters are employed in MFCC to weight the Discrete Fourier Transform (DFT) of the speech such that the outputs are approximately the energies of the filter-bank signals. The outputs of the band-pass filters are then used to calculate the MFCC using the Discrete Cosine Transform (DCT), which is defined as:

$$\mathrm{MFCC_i} = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos(\frac{\pi i}{N}(j - 0.5)) \quad (3.3)$$

Where N is the number of band-pass filter, m_j is the log of the j^{th} band-pass filter's output amplitude.

Each clip is a few seconds long and comes with its transcription. We then divide the speech into 25ms frames, each overlapping by 10ms, and multiply each frame by the Hamming window in order to make discontinuities smooth in the signal. From each of these frames, we extract features known as Mel-frequency Cepstral coefficients. The Mel-frequency Cepstrum is a representation of an audio signal on the Mel scale, a nonlinear mapping of frequencies, that down-samples higher frequency to imitate the human ear's ability to process sound. In our implementations, we used the first 13 Cepstral coefficients as our primary features, as is common in similar applications. A visualization of these features can be seen in figure 3.6.

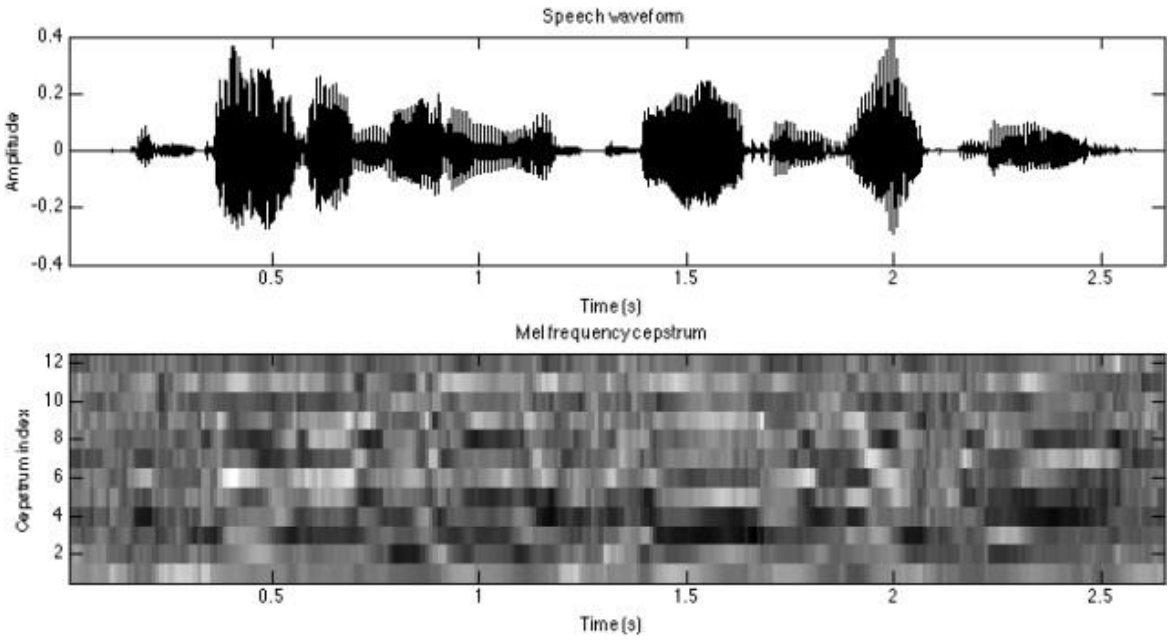

Figure 3.6 The original sound wave and its Cepstral coefficients

Taking the Cepstral coefficients as a base, we are using delta features. Delta features are the first and second time derivatives of the Cepstral coefficients, capturing the change of the Cepstral features over time, which we hypothesize, will be useful in classifying language, since pace is an important factor in language recognition by humans (Kim-Y. E. Wong, 2014). We can calculate these features as the central finite difference approximation of these derivatives

$$\Delta_{t,i} = \frac{c_{t+1,i} - c_{t-1,i}}{2} \quad (3.4)$$

And

$$\Delta_{t,i}^{2} = \frac{c_{t+1,i} - 2c_{t,i} + c_{t-1,i}}{4} \quad (3.5)$$

Where $c_{t,i}$; denotes the i^{th} coefficient of the t^{th} frame of the clip.

Power spectral estimation of a signal is one of the most important application areas in digital signal processing (1). Speech recognition techniques need spectrum analysis to measure speech bandwidth and further acoustic processing. Feature extraction makes an attempt to determine the quantities carrying information about the content of the speech and at the same time tries to eliminate irrelevant information (noise, distortion).

It creates series of feature vectors from the digitized speech signal. The role of the speech extractor unit is to extract the information from speech signal that is needed to identify the uttered word. Strive to filter out the factors that do not carry information about speech content. Hence, some transformation is needed to be done on the speech signal. It is usually performed in three main stages. The first stage is called the speech preprocessing. We designed the DNN architecture of the proposed system in the format as per expected requirement of the system and found from good result of experiments.

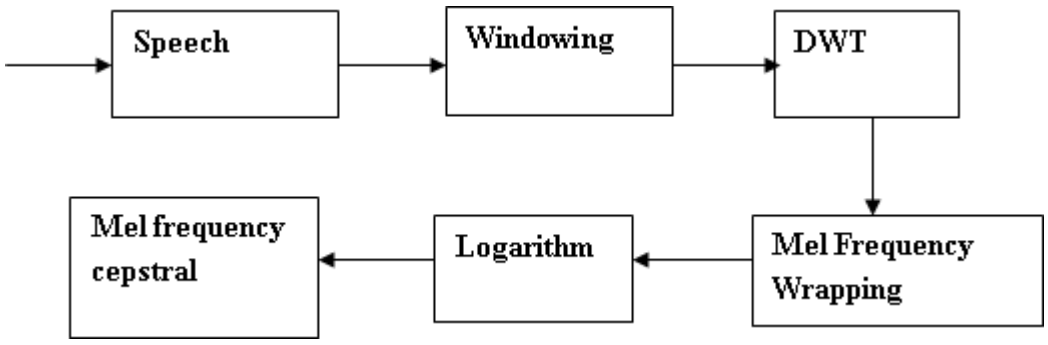

Figure 3.7 Feature Extraction.

CHAPTER FOUR

PROPOSSED SYSTEM

The proposed automatic speech recognition process consists of two major processes namely training and testing. In this proposed system the training is done by deep learning specifically fully connected deep recurrent neural network. In the front end we have employed the acoustic feature extraction Discrete Wavelet Transform and Mel-frequency Cepstral coefficient. As Figure 4.1 depicts that the task proceeding after collecting and arranging the sample data is preprocessing which is preparation of the speech samples for the feature extraction. Preprocessing involves silence removal, training and validation corpus preparation, comparing the audio with its corresponding text corpus. The unique features of human voice can be extracted by using Mel Frequency Cepstral Coefficient (MFCC) and this MFCC also represents the short term power spectrum of human voice. To calculate the coefficients which represent the frequency Cepstral, MFCC is used and these coefficients are based on the linear cosine transform of the log power spectrum on the nonlinear Mel scale of frequency. The final output of this system is predictive values of accuracy of training, validation and prediction, and their corresponding error values. We take the prediction accuracy and the word error rate as our evaluation metrics.

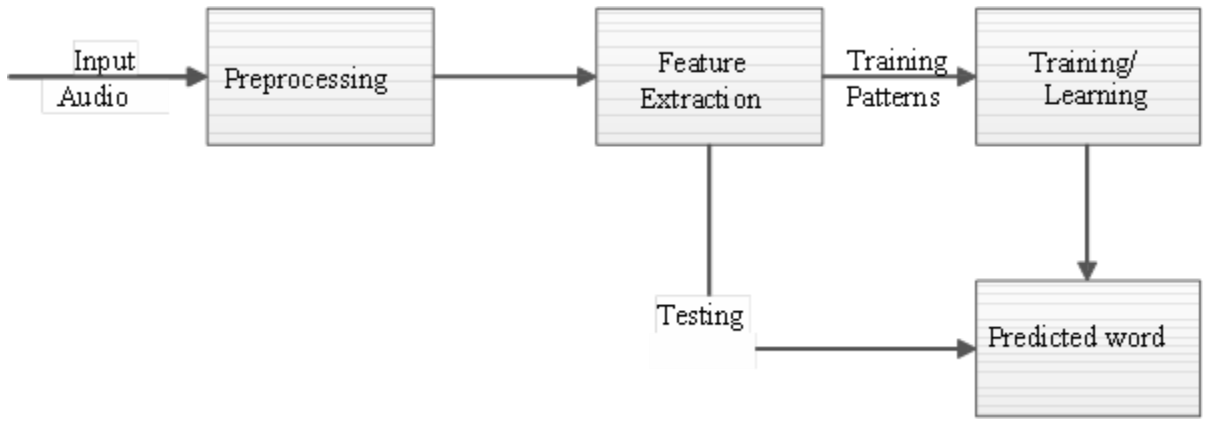

Figure 4.1 Proposed DRNN based ASR system block diagram

4.1 Training

Machine learning can be divided into supervised learning, unsupervised learning and semi-supervised learning based on its learning methods. Supervised learning, also known as learning with a teacher; which is different from unsupervised learning method by its assignment of labels for the samples in

the training dataset of the learning system. Details on supervised and unsupervised learning are described as follows.

4.1.1 Supervised Learning

Conceptually, we may think of the teacher as having knowledge about the environment, with that knowledge being represented by a set of input-output examples. However, the environment is unknown to the neural network. Suppose now that the teacher and the neural network are both exposed to a training vector (i.e., example) drawn from the same environment. By virtue of built-in knowledge, the teacher is able to provide the neural network with a desired response for that training vector. Indeed, the desired response represents the "optimum‖ action to be performed by the neural network. The network parameters are adjusted under the combined influence of the training vector and the error signal. The error signal is defined as the difference between the desired response and the actual response of the network. This adjustment is carried out iteratively in a step-by-step fashion with the aim of eventually making the neural network emulate the teacher; the emulation is presumed to be optimum in some statistical sense. In this way, knowledge of the environment available to the teacher is transferred to the neural network through training and stored in the form of ''fixed'' synaptic weights, representing long-term memory. When this condition is reached, we may then dispense with the teacher and let the neural network deal with the environment completely by itself (Simon Haykin, 2018).

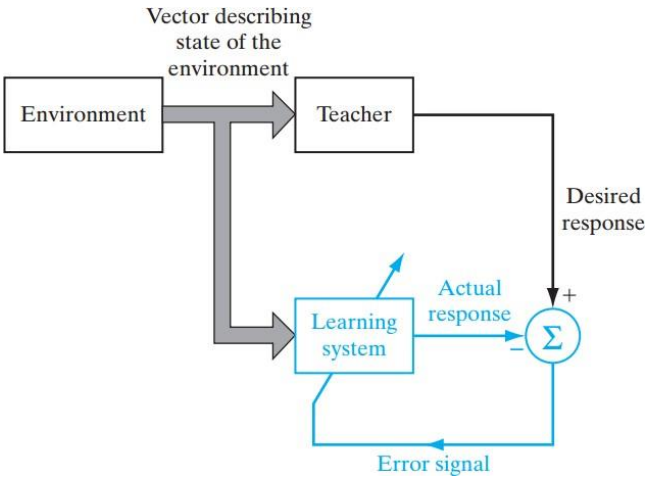

Figure 4.2 Block diagram of supervised learning (Simon Haykin, 2018)

The form of supervised learning we have just described is the basis of error correction learning. From figure 4.2, we see that the supervised-learning process constitutes a closed-loop feedback system, but the unknown environment is outside the loop.

As a performance measure for the system, we may think in terms of the mean square error or the sum of squared errors over the training sample, defined as a function of the free parameters (i.e., synaptic weights) of the system.

This function may be visualized as a multidimensional error-performance surface, or simply error surface, with the free parameters as coordinates. The true error surface is averaged over all possible input–output examples. Any given operation of the system under the teacher's supervision is represented as a point on the error surface. For the system to improve performance over time and therefore learn from the teacher, the operating point has to move down successively toward a minimum point of the error surface; the minimum point may be a local minimum or a global minimum. A supervised learning system is able to do this with the useful information it has about the gradient of the error surface corresponding to the current behavior of the system. The gradient of the error surface at any point is a vector that points in the direction of steepest descent. In fact, in the case of supervised learning from examples, the system may use an instantaneous estimate of the gradient vector, with the example indices presumed to be those of time. The use of such an estimate results in a motion of the operating point on the error surface that is typically in the form of a "random walk. Nevertheless, given an algorithm designed to minimize the cost function, an adequate set of input–output examples, and enough time in which to do the training, a supervised learning system is usually able to approximate an unknown input–output mapping reasonably well (Simon Haykin, 2018).

4.1.2 Unsupervised Learning/Learning without a Teacher

In supervised learning, the learning process takes place under the guidance of a teacher. However, in the paradigm known as learning without a teacher, as the name implies, there is no teacher to oversee the learning process. That is to say, there are no labeled examples of the function to be learned by the network (Becker, 1991, Simon Haykin, 2018). In unsupervised, or self-organized, learning, there is no external teacher or critic to oversee the learning process, as indicated in figure 4.3. Rather, provision is made for a task-independent measure of the quality of representation that the network is required to learn, and the free parameters of the network are optimized with respect to that measure. For a specific task-independent measure, once the network has become tuned to the statistical regularities of the input data, the network develops the ability to form internal representations for encoding features of the input and thereby to create new classes automatically (Becker, 1991).

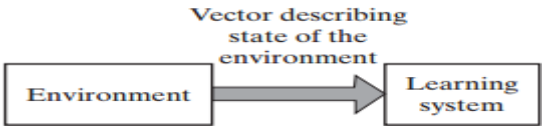

Figure 4.3 Block diagram of unsupervised learning

To perform unsupervised learning, we may use a competitive-learning rule. For example, we may use a neural network that consists of two layers an input layer and a competitive layer. The input layer receives the available data. The competitive layer consists of neurons that compete with each other (in accordance with a learning rule) for the ‘’opportunity’’ to respond to features contained in the input data. In its simplest form, the network operates in accordance with a ‘’winner-takes-all’’ strategy. In such a strategy, the neuron with the greatest total input ‘’wins’’ the competition and turns on; all the other neurons in the network then switch off (Simon Haykin, 2018).

4.1.3 Semi-supervised Learning

If there is large amount of the input data (X) and only some of the data is labeled (Y) are called semi-supervised learning problems. These problems sit in between both supervised and unsupervised learning. A good example is a photo archive where only some of the images are labeled, (e.g. dog, cat, person) and the majority are unlabeled. Many real world machine learning problems fall into this method. This is because it can be expensive or time-consuming to label data as it may require access to domain experts. Whereas unlabeled data is cheap and easy to collect and store (Simon Haykin, 2018). Unsupervised learning techniques are used to discover and learn the structure in the input variables. They are also used to make best guess predictions for the unlabeled data, feed that data back into the supervised learning algorithm as training data and use the model to make predictions on new unseen data.

4.2 Modeling

Automatic speech recognition (ASR) refers to the task of recognizing human speech and translating it into text. This research field has gained a lot of focus over the last decades. It is an important research area for human-to-machine communication. Early methods focused on manual feature extraction and conventional techniques such as Gaussian Mixture Models (GMM), the Dynamic Time Warping (DTW) algorithm and Hidden Markov Models (HMM). More recently, neural networks such as recurrent neural networks (RNNs), convolutional neural networks (CNNs) have been applied on ASR and have achieved great performance.

4.2.1 Deep Learning and Recurrent Neural Networks

Deep learning is a subset of a broader family of machine learning methods based on learning data representations, which is contrary to task-specific algorithms as they are not learning based.

The difference between shallow and deep learning is the depth of their credit assignment paths, which is chains of possibly learnable, causal links between actions and effects. This has been called the fundamental credit assignment problem concerns about which modifiable components of a learning system are responsible for its success or failure and what changes to them improve performance. There are general credit assignment methods for universal problem solvers that are time-optimal in various theoretical senses. A standard neural network (NN) consists of many simple, connected processors called neurons, each producing a sequence of real-valued activations. Input neurons get activated through sensors perceiving the environment; other neurons get activated through weighted connections from previously active neurons. Some neurons may influence the environment by triggering actions. Learning or credit assignment is about finding weights that make the NN exhibit desired behavior. Depending on the problem and how the neurons are connected, such behavior may require long causal chains of computational stages, where each stage transforms (often in a non-linear way) the aggregate activation of the network. Deep learning is about accurately assigning credit across many such stages. RNNs are the state of the art algorithm for sequential data. Recurrent Neural Networks are well suited for tasks that involve sequential data. And because speech is sequential data, we could train a RNN with a dataset of acoustic observations and their corresponding transcriptions.

4.2.2 Deep Learning Based Speech Recognition

The strength of neural network based solutions lies in the fact that they can be used as generic learning methods to any problem. Many research papers have employed deep learning in the Automatic Speech Recognition process either as front-end which can effectively mine the contextual information embedded in speech frames or back-end utilizing large amounts of labeled training data (https://www.ethiopiaonlinevisa.com/amharic-the-ethiopian-language/, I. L. Moreno, J.G. Dominguez, D. Martinez, O. Plchot, J.G. Rodriguez, P.J. Moreno, 2014). In this thesis work we used Deep Recurrent Neural Networks (DRNN). Because it has shown excellent results and tech companies including Google, Amazon and Baidu use DL for speech recognition. The core idea is to have a network of interconnected nodes (Neural Networks) where each node computes a function and passes information to the nodes next to it. Every node computes its output using the function it was configured

to use, and some parameters. The process of learning updates these parameters. The function itself will not change. Initially the parameters are randomly initialized. Then the training data is passed through the network, making small improvements to the parameters on every step.

4.3 Acoustic Model

The development of any speech recognition system consists the following steps for the construction of ASR system. The steps are. Pre-processing, feature extraction, decoding (using acoustic and language model). An acoustic model (AM) is used in automatic speech recognition to represent the relationship between an audio signal and the phonemes or other linguistic units that make up speech. The model is learned from a set of audio recordings and their corresponding transcripts. It is created by taking audio recordings of speech, and their text transcriptions. Modern speech recognition systems use both an acoustic model and a language model to represent the statistical properties of speech. The acoustic model models the relationship between the audio signal and the phonetic units in the language. The language model is responsible for modelling the word sequences in the language. These two models are combined to get the top-ranked word sequences corresponding to a given audio segment. The acoustic models are statistical models which capture the correspondence between a short sequence of acoustic vectors and an elementary unit of speech. The elementary units of speech that are most often used in ASR are phones. Phones are the minimal units of speech that are part of the sound system of a language, which serve to distinguish one word from another. Most modern speech recognition systems operate on the audio in small chunks known as frames with an approximate duration of 10ms per frame. The raw audio signal from each frame can be transformed by applying the Mel-frequency Cepstrum.

The coefficients from this transformation are commonly known as Mel frequency Cepstral coefficients (MFCCs) and are used as an input to the acoustic model along with other features. We have pre-processed audio signals to get MFCC features, and worked out how network predictions can be decoded into phones. We need to associate audio signals, represented by MFCC features, with Amharic characters. This part is known as acoustic modeling.

4.5 Language Model

Language model is one of the most important knowledge sources for large vocabulary Speech Recognition System. The language model contains rudimentary syntactic information. Its aim is to predict the likelihood of specific words occurring one after another in a given language. The main source for training a LM for conversational speech are the transcriptions of the audio training corpora. LM in speech recognition would have a tighter integration with the network. The network would predict probabilities based on speech, and the LM would help predict words. LM requires much larger training data (audio with its transcription) because we would be training the model not only as to how to translate speech utterances into characters, but also how people normally fit words together in that language. So it's easier to train on textual data separately and then use trained LM model to aid the acoustic model.

4.6 Power Spectrum Density Estimation

In statistical signal processing, the goal of spectral density estimation (SDE) is to estimate the spectral density (also known as the power spectral density) of a random signal from a sequence of time samples of the signal. One of the very paramount application zones of digital signal processing is the power spectrum estimation of periodic and random signals. Speech recognition issues utilize spectrum analysis as a preliminary measurement to perform speech bandwidth reduction and further acoustic processing (Maazouzi, A. Laaroussi, N. Aqili, M. Raji and A. Hammouch LRGE, ENSET, Mohammed, 2016). The discrete wavelet transforms with discrete time Cj, m[k] will be calculated according to the relation.

$$\text{Cj, m[K]} = \sum_{K=0}^{N-1} \text{f(In Pxx [K]} - \text{y)†j, m[}K\text{]} \quad (4.1)$$

4.7 Discrete Wavelet Transform

The Fourier-based transforms such as the Discrete Fourier Transform (DFT), DCT, and MDCT have a problem encoding highly non-stationary signals because the transform of a non-stationary signal spreads over the whole spectrum, which makes compression in the transform domain a more difficult task (Michael MeleseWoldeyohannis, 2016). Wavelet transform decomposes a signal into a set of basis functions. These basis functions are called wavelets. Discrete wavelet transform is a widely used method in many different fields such as signal processing, biomedical signal processing,

communication in recent years for feature extraction (Kim, Euisung, 2016). The Discrete Wavelet Transform (DWT) is the total of the signal that is scaled over the time domain of the wavelet over time and multiplied by its offset values (Kim, Euisung, 2016).

In wavelet analysis, the Discrete Wavelet Transform (DWT) decomposes a signal into a set of mutually orthogonal wavelet basis functions.

These functions differ from sinusoidal basis functions in that they are spatially localized – that is, nonzero over only part of the total signal length. The DWT, in fact, refers not just to a single transform, but rather a set of transforms, each with a different set of wavelet basis functions. Two of the most common are the Haar wavelets and the Daubechies set of wavelets. The signal decomposition called the discrete wavelet transform (DWT) and the DWT coefficients are formed by the high-frequency coefficients of each level together with the low-frequency coefficients of the last level (Michael MeleseWoldeyohannis, 2016) Because of the decimation by 2 at each level, the DWT of an *L* dimensional signal will produce *L* transform coefficients. Discrete Wavelet Transform is a technique to image compression, advanced signal processing (speech and image processing), into wavelets, which are then used for wavelet-based compression and coding. The DWT is defined as [Gonzalez]:

$$W\emptyset(J_0, K) = \frac{1}{\sqrt{M}} \sum_x f(x) \emptyset J_0, K(x) \quad (4.2)$$

$$W\varphi(J, K) = \frac{1}{\sqrt{M}} \sum_x f(x) \varphi J, k(x) \quad (4.3)$$

For j ≥ j0 and the Inverse DWT (IDWT) is defined as

$$f(x) = \frac{1}{\sqrt{M}} \sum_x W\emptyset(J_0, K) \emptyset J_0, K(x) + \frac{1}{\sqrt{M}} \sum_{J-j0}^{\infty} \sum_x W\varphi(J, k) \emptyset J, k(x) \quad (4.4)$$

Where f(x),$\emptyset J_0, K(x)$and φJ, k(x) are functions of the discrete variable.

The $\emptyset J_0, K(x)$is a member of the set of expansion functions derived from a scaling function ∅(x), by translation and scaling using

$$\emptyset J, k(x) = 2^{J/2} \emptyset(2^j x - k) \quad (4.5)$$

The $\emptyset J, k(x)$ is a member of the set of wavelets derived from a wavelet functionφ(x) by translation and scaling using:

$$\varphi j, k(X) = 2^{j/2} \varphi(2^j X - K) \quad (4.6)$$

The methodology can be summarized into specific components or modules as following

I. Dataset collection: collecting data to prepare the dataset.
II. Pre-processing: The speech sections must be separated from non-speech to reach optimal recognition accuracy in a practical environment. Separating the speech
III. Features Extraction: using MFCC and DWT.
IV. Modeling: implement the proposed system with python
V. Result and Discussion: Analysis and discussion on the results, problems faced recommendation and future works.

Tools: Feature extraction training and testing processes is done by **Python,** audacity to record audios and different downloading tools to download files

CHAPTER FIVE

IMPLEMENTATION OF THE SYSTEM

5.1 Experiments

The aim of these experiments is to show that Amharic Language can be used to develop Automatic Speech Recognition System. It can be also developed to an application level by using neural network with different feature extraction methods like MFCC and DWT and using tools for graphical user interface like python.

5.2 Data Processing

We obtained our training and testing data from ALFFA_PUBLIC (African Languages in the Field, Speech Fundamentals and Automation (http://alffa.imag.fr/)) which is working on Kaldi Speech Recognition Tool, by recording different speakers, and downloading videos from YouTube. Providing a collection of uploaded video. We implemented a signal processing to remove silence. The data recorded from different speakers (male and female) it is around 20 hours of data. The total number of samples used are more than 8000 for training and testing. Each clip is a few seconds long (2 to 18 words) and comes with its transcription.

5.3 System Settings

In this work we chose to use DWT and MFCC over Spectrogram because it produces fewer numbers of features. Let's take a few seconds of audio clip and plot its MFCC features. The result is shown below. Time is on the Y axis. Each horizontal line represents a short audio clip of 10ms duration. The feature vectors used comprised 13 order Mel-frequency Cepstral coefficients (MFCCs) derived from 20 filter banks.

Each feature vector is extracted at 10ms intervals using a 32ms window of band limited (300 - 3400 Hz) speech. Each frame of speech is pre-emphasized by the first order difference equation

$$s'_n = s_n - 0.97 s_{n-1} \quad (5.1)$$

Where s'_n new n^{th} frame of speech is, s_n is the old n^{th} frame of speech and s_{n-1} is the speech frame which is previous of n^{th} frame. A Hamming window was then applied to the speech frame. The 13

columns (0-12) represent what's known as MFCC coefficients in signal processing.

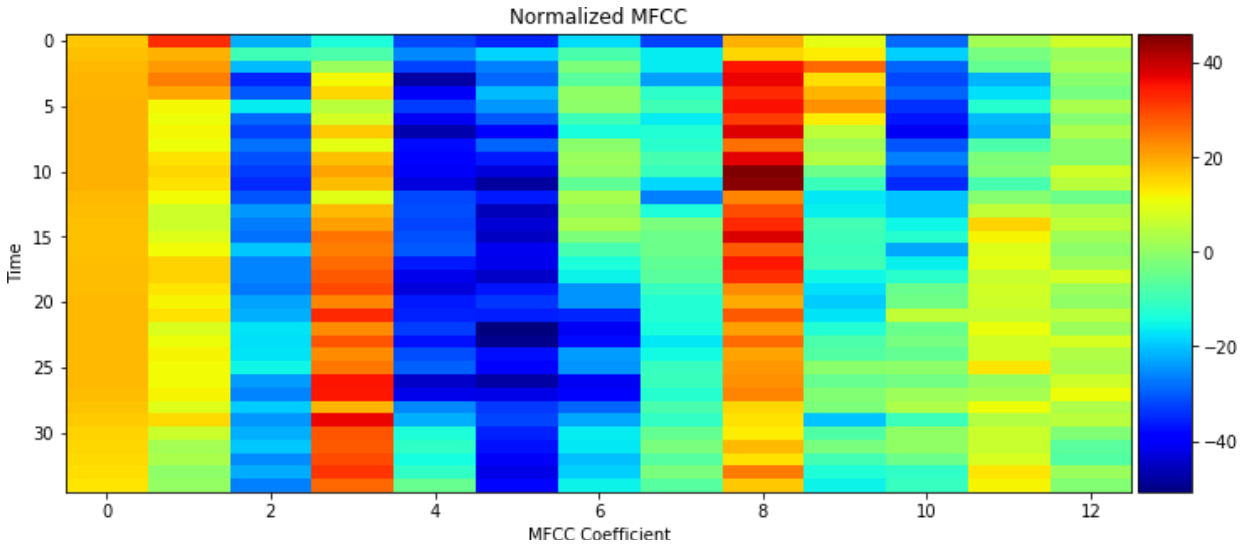

Figure 5.1 MFCC coefficient

Wavelet transform is applied to evaluate it's time frequency relation

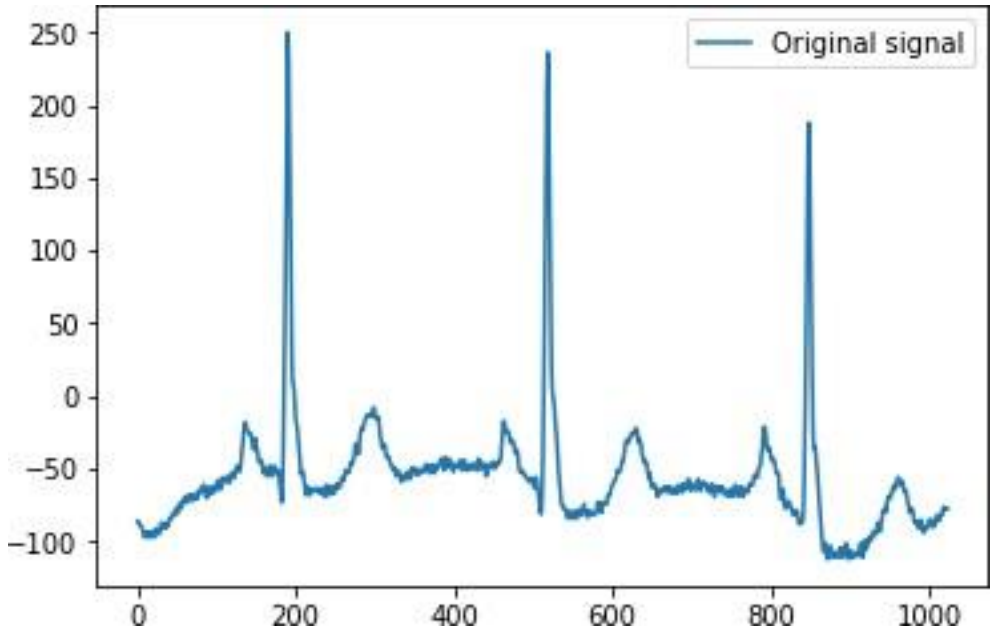

Figure 5.2 Discrete Wavelet transforms

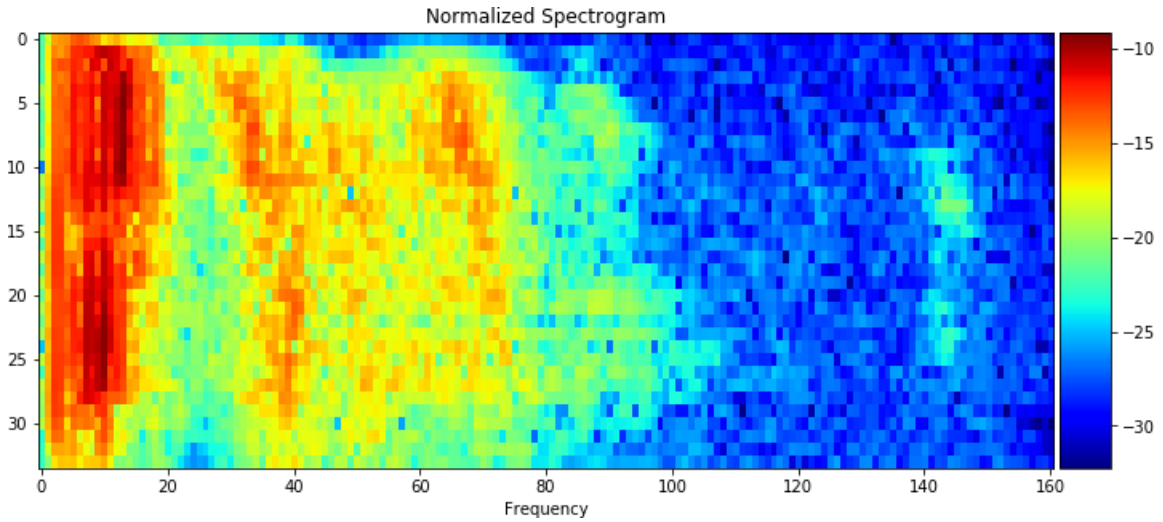

Figure 5.3 Normalized Spectrogram

5.4 Evaluation

The Deep Learning based speech recognition system can depend on acoustic speech features. Thus the acoustic feature vectors employed directly influence the accuracy of the system. The most common Evaluation metrics used for Speech Recognition System is Word Error Rate (WER).

Word Error Rate

The Word-Error-Rate WER operates on the word-level and is a common metric for Automatic Speech Recognition or Machine Translation models. WER is a minimum edit-distance measure produced by applying a dynamic alignment between the output of the ASR system and a reference transcript. In the alignment process, three types of errors can be distinguished:

- A **substitution** occurs when a word gets replaced (for example, in the sentence "ያንደኛ ደረጃ ትምህርታቸውን ጐንደር ተምረዋል'' the word ጐንደር transcribed as ቆንዳር)
- An **insertion** is when a word is added that wasn't said (for example, in the sentence "አስር ሺ ኢትዮጵያዊያንም ቤት አልባ ሆኑ'' the word ''ቤት'' is added and transcribed as ''አስር ሺ ኢትዮጵያዊያንም ቤት አልባ ቤት ሆኑ'')
- A **deletion** happens when a word is left out of the transcript completely (for example, in the sentence "ሰውዬው ሚስቱን በጣም እንደሚወዳት በለሆሳ ስነገራት'' the word ''በለሆሳ'' is deleted)

- **substitution**-errors concern words which were transcribed wrongly (Word-A which is in

the reference sentence was transcribed as Word-B or wrong word which is not in the reference sentence), whereas **insertions** and **deletions** concern words which were transcribed in addition to reference words or words omitted during transcription.

In order to calculate the WER, the recognizer is used to transcribe audio corresponding to the test set. The resulting transcript is then compared to the true transcription provided as reference.

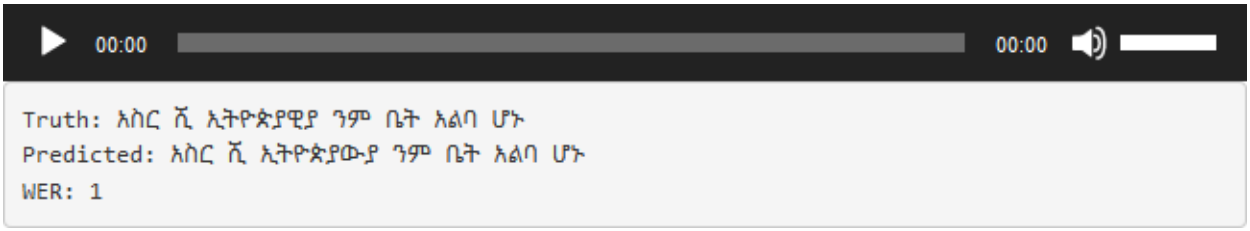

Figure 5.4 WER

WER can be calculated as follows.

$$\text{WER} = \frac{100 * (\text{sub} + \text{ins} + \text{del})}{\text{total number of words in reference transcript}}$$

In the sentence አስር ሺ ኢትዮጵያዊያንም ቤት አልባ ሆኑ

The predicted sentence is አስር ሺ ኢትዮጵያውያንም ቤት አልባ ሆኑ there is one-word substitution

Example to show WER. $\text{WER} = \frac{100*(1+0+0)}{6} = 16.6\%$

This result show that 84.3 % accuracy of the sentence.

Table 5.1 Sample Training WER

Sample Sentence		
Truth	Predicted	WER
የሚወጡት ኤምባሲው ውስጥ የሚሰሩ አባላትና ጥገኞቻቸው እንደሆኑ አሜሪካኖቹ አስታውቀዋል	የሚጡት ምባሲ ውስጥ የሚሰር አባላት ነው ትገንቻቸው እንዱሙ አሜሊካንቹ አስታውዋል	7
አስር ሺ ኢትዮጵያዊያንም ቤት አልባ ሆኑ	አስር ሺ ኢትዮጵያውያንም ቤት አልባ ሆኑ	1
ሰውዬው ሚስቱን በጣም እንደሚወዳት በለሆሳስ ነገራት	ሰውዬው ሚስቱን በጣም እንደሚወዳት በለሆሳስ ነገራት	0
ያንደኛ ደረጃ ትምህርታቸው ን ጐንደር ተምረዋል	ያንደኛ ደ- ጃ ትም-ርታቸው ቆንዳር ተምረዋል	3

```
processed 10 in 1 minutes
processed 20 in 2 minutes
processed 30 in 3 minutes
processed 40 in 4 minutes
processed 50 in 5 minutes
processed 60 in 6 minutes
processed 70 in 7 minutes
processed 80 in 8 minutes
processed 90 in 9 minutes
processed 100 in 10 minutes
processed 110 in 11 minutes
processed 120 in 12 minutes
processed 130 in 13 minutes
processed 140 in 14 minutes
processed 150 in 15 minutes
processed 160 in 16 minutes
processed 170 in 17 minutes
processed 180 in 18 minutes
processed 190 in 19 minutes
processed 200 in 20 minutes
```

Figure 5.5 WER processing time

Results on Training the System

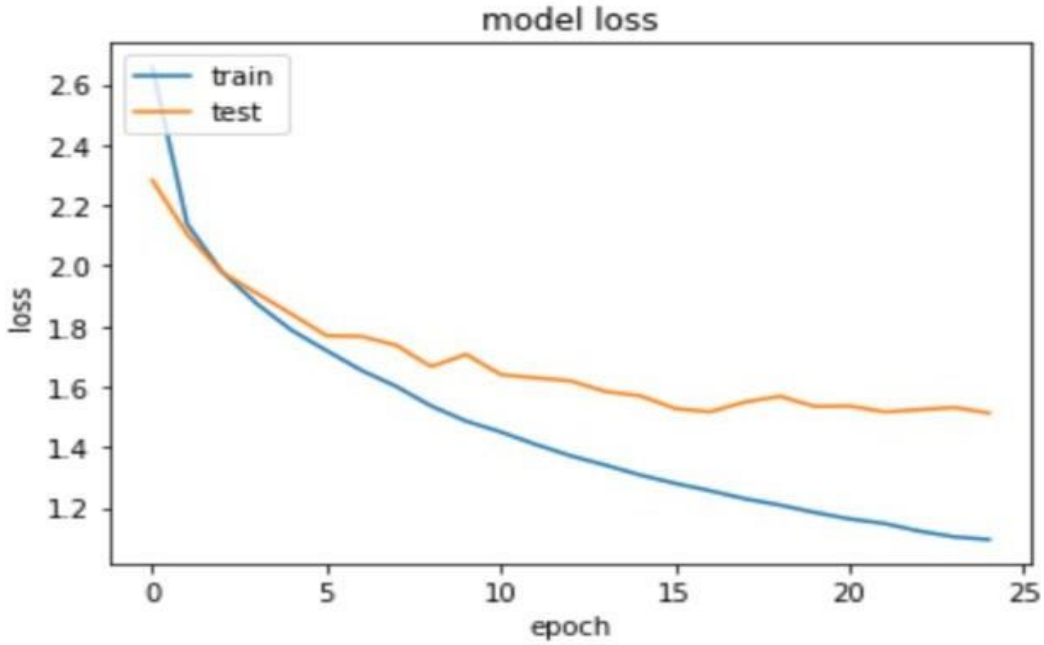

Figure 5.6 The Training loss of system

As shown in the above graph the training loss will decreases through the training increases

Prediction Accuracy of the System

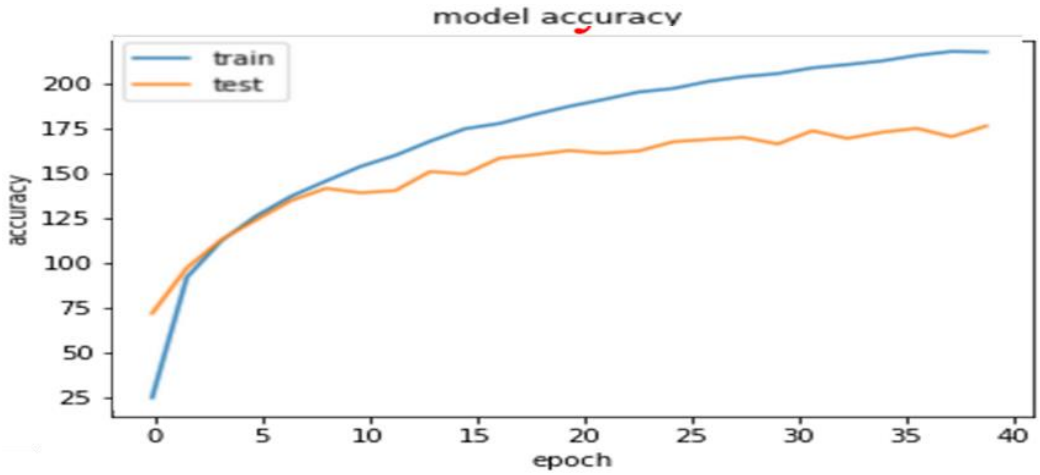

Figure 5.7 Prediction accuracy of the system.

The recognition accuracy increases as the training processes increases

Real Time Test.

The accuracy of the real time test is lower than the actual validation because of environmental factors like background nose, pronunciation, behavior of the speaker. As shown in the figure below there is substitution of words which are not spoken. WER in validation is 7.02 or 92.98 % accurate but at the time of real time test the WER is 19.8 or 80.2% accuracy.

ዋናው
አዲሱ የኢትዮጵያ ብሄራዊ ቡድን አሰልጣኝ ስራቸውን ጀመሩ
አዲሱ የኢትዮጵያ ብሄራዊ ቡድን አሰልጣኝ ስራቸውን ጀመሩ
የተለቀቁት መግለጫ እኮ
የኢትዮጵያ በላይ የኤርትራ አመልክ ያቀበለዋል ሊቀመንበር ከለቦች ነው
እስካሁን ከግሉ የፖሊሲ ያስፈልጋል አቀርባለሁ በመግባት ትላንት መውሰድ
በአለም ጉባኤ አድርገው
በእርዳታው
በአለም ጉባኤ አይጠበቅም
እንዲሳተፍ ሀይሎች ናቸው
የተለቀቁት ልማዳቸው

Figure 5.8 Real time Test.

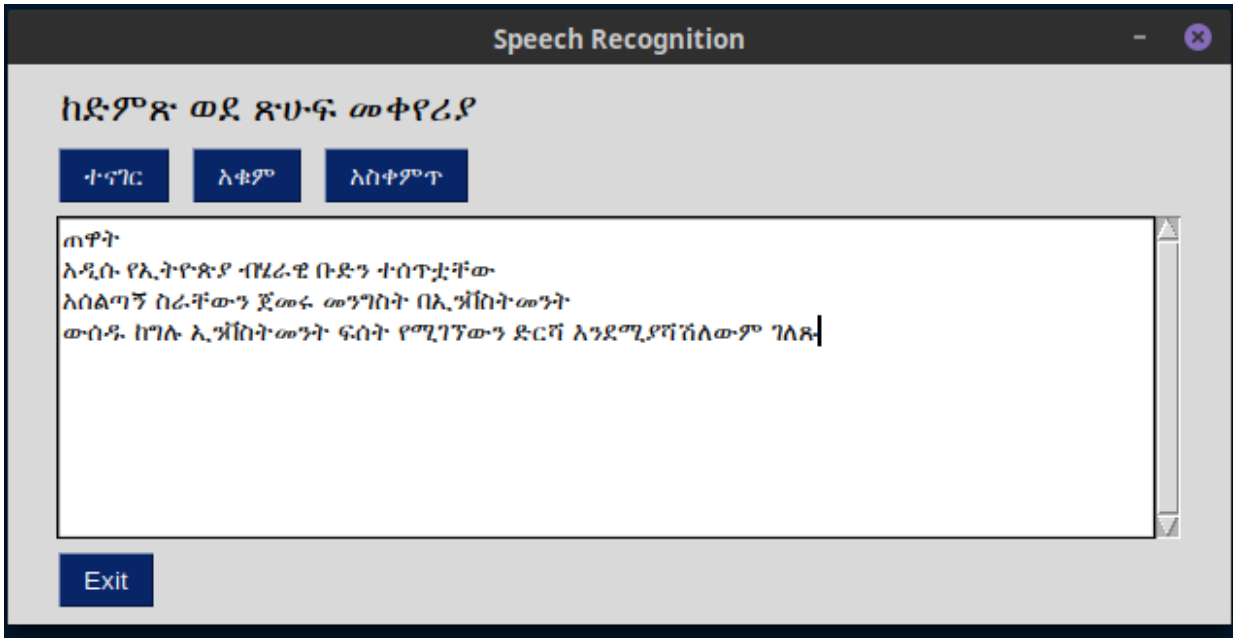

Figure 5.9 Real Time Test with GUI

The real time test with simple GUI is to show that the system can be developed as an application for computers and other devices. The accuracy of this application is with WER 19.8.

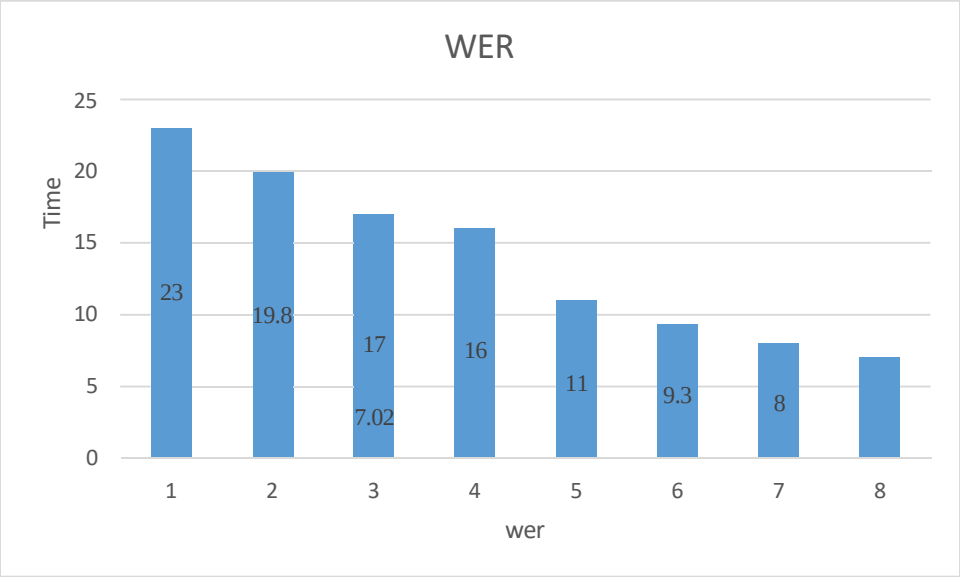

Figure 5.10 WER accuracy graph

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

Speech recognition is a field which has been and being researched for more than a decade for most of the resourced languages like English and most European languages. On the other hand, attempts in this area for under resourced languages like Amharic speech recognition, in particular, is at early level. The objective of this work is to implement Speaker Independent and continuous Amharic Language Speech Recognition System. In this thesis work, 20 hr. audio with its transcription is prepared to facilitate Amharic speech recognition which is from (http://alffa.imag.fr/) with 16hr audio and the rest 4hr. audio is collected from different medias, and recorded speakers.

We have achieved the main objective of our work, which is exploration of possibilities for the development of speaker independent and continuous speech recognition for Amharic. As we tried to implement continues speech recognition the result shows satisfying accuracy which is 19.8 WER or 80.2% accuracy in real time and best result when validation audio testing which is 7.02 WER or 92.98% accuracy. To meet the objective of this work Sample audios are collected and transcribed to text. The audios are recorded from different speakers. The corpus we have developed and used to develop the ASR for Amharic is medium sized Amharic read speech corpus. Features are extracted for the training purpose using discrete wavelet transform and Mel frequency Cepstral coefficient. The system is tested using validation audio and at real time speech. Using validation audio, the system predicts the words or sentence with minimum word error rate. The word error rate shows minimum increase in real time test, because of environmental noise, and other factors.

6.2 Recommendations

We have achieved significant progress in the development of speaker independent, continuous speech recognition for Amharic. However, we see that there is a lot of work remaining in the area. To get Data for Amharic speech recognition is extremely difficult. For this thesis work we use 10k or about 20hr. audio data with its transcription. The performance of the continuous, speaker independent Amharic Speech Recognition System can further be enhanced by adding pronunciation variants to all words that have pronunciation variants. Preparing more data at different environment, different age

range of speakers, with multiple pronunciation variation, with noisy environment is crucial. Beside this, to get better accuracy, using different deep learning methods and feature extraction methods is advisable. Improving the Acoustic model by using hybrid acoustic modeling using LSTM and other acoustic modeling algorithms. The training also needs high performing hardware specifications. Using GPU system is more advisable to minimize the training time.

REFERENCES

Maazouzi, A. Laaroussi, N. Aqili, M. Raji and A. Hammouch LRGE, ENSET, Mohammed V University in Rabat, Morocco ENSEM, Hassan II University in Casablanca, Morocco (2016), ‘‘Speech Recognition System using Burg Method and Discrete Wavelet Transform’’.

Kiran.R, Nivedha.K, Subha.T and Pavithra Devis Department of information Technology, Sri Sai Ram engineering collage, Chennai-44, Tamil Nadu (2017) ‘‘Voice and speech recognition in Tamil language’’.

Basem H. A. Ahmed and Ayman S.Ghabayen, Department of computer science and Technology University collage of Science and Technology Gaza, Palestine (2017) ‘‘Arabic Automatic Speech Recognition Enhancement’’.

Abraham Woubie Zewoudie Addis Ababa University school of graduate studies (2011) ‘‘Enhanced Amharic speech recognition systems’’.

Suma Swamy1 and K.V Ramakrishnan1Department of Electronics and Communication Engineering, Research Scholar, Anna University, Chennai (2013) ‘‘An efficient speech recognition system’’.

Kim, Euisung, (2016) ‘‘A Wavelet Transform Module for a Speech Recognition Virtual Machine’’.

Michael MeleseWoldeyohannis, Laurent Besacier , Million Meshesha 1, Addis Ababa University, Addis Ababa, Ethiopia 2, LIG Laboratory, UJF, BP53, 38041 Grenoble Cedex 9, France (2016) ‘‘Amharic Speech Recognition for Speech Translation’’.

SevcanAytaçKorkmaz and Furkan Esmeray Electric and Energy Department Munzur University Tunceli, Turkey (2018) ‘‘Quality Lignite Coal Detection with Discrete Wavelet Transform, Discrete Fourier Transform, and ANN based on k-means Clustering Method’’.

TokunboOgunfunmi and Koji Seto Dept. of Electrical Engineering Santa Clara University Santa Clara, California, CA 95053, USA (2016). ‘‘On the Use of Discrete Wavelet Transform for Robust Scalable Speech Coding’’.

Abd-ErrahimMaazouzi*, Nabil Aqili, MouradRaji, Ahmed Hammouch LRGE Laboratory, ENSET University of Mohammed V Rabat, Morocco (2019) ‘‘A Speaker Recognition System Using Power Spectrum Density and similarity measurements’’.

Arkady Produs, Kateryna Kukharicheva, (2017) ‘‘Automatic Speech Recognition Performance For Training on Noised Speech’’.

I. L. Moreno, J.G. Dominguez, D. Martinez, O. Plchot, J.G. Rodriguez, P.J. Moreno, (2014)

‘‘Automatic Language Identification Using Deep Neural Networks’’

Https://En.Wikipedia.Org/Wiki/Deep_Learning

Kim-Y. E. Wong, ‘‘Automatic Spoken Language Identification Utilizing Acoustic And Phonetic Speech Information’’ Phd Thesis, June 2004.

Y. Song, B. Jiang, Y. Bao, Si Wei And Li-R. Dai, (2013) ‘‘i-Vector Representation Based On Bottleneck Features For Language Identification’’

Simon Haykin, Book 2008 ‘‘Neural Networks and Learning Machines’’.

https://www.ethiopiaonlinevisa.com/amharic-the-ethiopian-language/

S. Ranjan, C. Yu, C. Zhang, F. Kelly, J. H. L. Hansen, 2016) ‘‘Language Recognition Using Deep Neural Networks With Very Limited Training Data’’.

Ankita Chugh, Poonam Rana, Suraj Rana (2014)‘‘Speech Recognition System Using Wavelet Transform’’.

Vimal krishinan VR, Athulya Jayakumar, Babu Anto.P (2018) “ Speech Recognition of Isolated Malayalam Words Using Wavelet Features and Artificial Neural Network”.

Tanja Schultz the 9th International symposium on chines spoken language processing (2014) “Multilingual Automatic Speech Recognition for Code-switching Speech”.

Martha Yifru Tachbelie, Solomon Teferra Abate, Wolfgang Menzel “Morpheme-based Automatic Speech Recognition for a Morphologically rich language – Amharic’’.

Asadullahi, Arslan Shaukati, Hazrat Ali, Usman Akrami (2016) ‘‘Automatic Urdu Speech Recognition Using Hidden Markov Mode’’

Dr.E.Chandra, A.Akila (2012) ‘‘An Overview of Speech Recognition and Speech Synthesis Algorithms”.

Rainer Martin Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics”

Chadawan Ittichaichareon, Siwat Suksri and Thaweesak Yingthawornsuk “Speech Recognition Using MFCC”.

George Tzanetakis, Georg Essl, Perry Cook ‘‘Audio Analysis using the Discrete Wavelet Transform”

Mehryar Mohri Michael Riley Richard Sproat Tutorial presented at COLING 96, August 3rd, 1996 “Algorithms for Speech Recognition and Language Processing”.

Tufekci, J.N. Gowdy IEEE 2000 “FEATURE EXTRACTION USING DISCRETE WAVELET TRANSFORM FOR SPEECH RECOGNITION”

Tingxiao Yang (2012) ‘‘The Algorithms of Speech Recognition, Programming and Simulating in MATLAB’’

Marek Gliszczynski_, Galina Cariowa_, Aleksandr Cariow IEEE 2013 “An algorithm for DWT basic procedure rationalized implementation”

Irina Kipyatkova Alexey Karpov “Recurrent Neural Network-based Language Modelling for an Automatic Russian Speech Recognition System”

Solomon Teferra Abate, Wolfgang Menzel, Bairu Tafila,‘‘An Amharic Speech Recognition Corpus for Large Vocabulary Continuous Speech Recognition” Fachbereich Informatik, Universit¨at Hamburg

Hamdan Prakoso1, Ridi Ferdiana, Rudy Hartanto ISESD (2016) “Indonesian Automatic Speech Recognition System Using CMUSphinx Toolkit and Limited Dataset”.

Pavel Kr´al University of West Bohemia “Discrete Wavelet Transform for Automatic Speaker Recognition”

Pavel Krbec “Language Modelling for Speech Recognition of Czech”.

Md. Akmal Haidar (2014) ‘‘Language Modelling for Speech Recognition Incorporating Probabilistic Topic Models”

Jungsuk Kim and Wonyong Sung IEEE 2012 “MULTI-USER REAL-TIME SPEECH RECOGNITIONWITHA GPU”

https://en.wikipedia.org/wiki/Speech_recognition

Christoph Bensch (2021) ‘‘Continuous Learning in Automatic Speech Recognition’’