ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

SPEAKER DEPENDENT SPEECH RECOGNITION FOR SIDAAMA LANGUAGE

By

ABDELLA KEMAL MOHAMMED

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

July, 2010

Addis Ababa,Ethiopia

ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

SPEAKER DEPENDENT SPEECH RECOGNITION FOR SIDAAMA LANGUAGE

BY

ABDELLA KEMAL MOHAMMED

Signature of Board of Examiners for Approval

DEREJE TEFERI (Ph.D.)	_____
Advisor	Signature
_____	_____
Examiner	Signature
_____	_____
Chairman	Signature

DEDICATION

To my parents. Love you All.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude and appreciation to Dr Dereje Teferi, my advisor, for his guidance, valuable suggestions and encouragement. This thesis would never have been possible without his insightful observations and advice. It has been a privilege and pleasure to have worked with him.

It is also my pleasure to acknowledge the Faculty of Informatics; Particularly the Department of Information Science for financing the research project for its successful completion, with out it would have been very hard.

It is also my pleasure to acknowledge Hawassa College of Teacher Education (HCTE) for the support and encouragement I had during my study. I also acknowledge Girum Tesfaye, who is a student in department of Linguistics, AAU for his help and valuable suggestions during the selection and recording of Sidaama words for the study.

I am also grateful for the assistance I always had from my colleague and best friend Abraham Lule, assistant Lecture in department of ICT, HCTE.

Finally, I would like to acknowledge my mom and dad, my fiancé, and my brothers and sisters for all the encouragement, support and love they gave me during the study. It was a strengthening energy for successful completion of my work.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
CD	Context Dependent
CI	Context Independent
CMU	Carnegie Mellon University
DARPA	Defence Advance Research Projects Agency
DTW	Dynamic Time Warping
GB	Giga Byte
HEC	Highland East Cushitic
HMM	Hidden Markov Model
IBM	International Business Machine
ICT	Information and Communication Technology
IPA	International Phonetic Alphabet
LPA	Linear Predictive Analysis
LPC	Linear Predictive Coding
MB	Mega Byte
MERL	Mitsubishi Electric Research Laboratory
MFCC	Mel-Frequency Cepstral Coefficient
PC	Personal Computer
PLP	Perceptual Linear Prediction
RAM	Random Access Memory
TGC	Transitional Government of Ethiopia
WER	Word Error Rate

ABSTRACT

Speech recognition systems have been applicable in wide areas as various speech recognition methodologies, techniques and tools have been developed and implemented to generate a natural and intelligible speech. The main objective of this thesis is to explore the possibility of developing prototype speaker dependent speech recognition for Sidaama language using Hidden Markov Model(HMM). In order to come up with a working prototype model for the language extensive study was conducted on the language to understand and come up with the language features needed to build the model.

Additionally, the components as well as techniques used in the HMM based speech recognition design were studied and analyzed to identify those components that are dependent on the characteristics of the language. Besides the most commonly used speech recognition tools were critically reviewed and as a result the most widely used Java based speech recognizer tool called the Sphinx Systems was used to build the acoustic models as well as for testing the recognition performance.

This research attempted to build context dependent triphone based isolated speech recognizer as well as context independent monophone based isolated word speech recognizer models for Sidaama language. A total of 450 unique words were selected and recorded in consultation with a domain linguistic expert. Out of the total datasets, 300 of the recorded words were used for training the acoustic models whereas the remaining 150 words were used for testing the performances of the constructed acoustic models. In addition out of the 300 words used for training the HMM acoustic model 100 words were randomly selected and used for testing constructed models.

The performance of the context dependent triphone based model achieved 73% accuracy for 100 words selected among 300 words used for building the acoustic models whereas 68% accuracy is obtained using 150 words which were not included in building the recognizer model. Similarly, the context independent word based model achieved 69% accuracy for 100 words selected among the 300 words used for building the context dependent acoustic model where as 58% accuracy was achieved using 150 words which were not used for building the acoustic model. As a result the context dependent triphone based model is suggested to be appropriate for building speech recognizer for Sidaama language. In conclusion the results obtained were encouraging and more optimization works should be done in the future to improve the recognition performance.

Keywords: Automatic Speech Recognition, Sidaama Language, Sphinx System, HMM

CHAPTER ONE

INTRODUCTION

In this chapter, an overview of speech recognition systems is briefly discussed and the problem statement and justification that initiated the research in the first place are discussed. Furthermore, the general and specific objectives of the study along with the methods used to achieve the intended outcome are put in place. Finally, the scopes and the limitations of the study, the possible applications sought and the organization of the thesis are presented.

1.1. Background

Speech is the most natural mode of communication between people. People learn the relevant skills during early childhood without instruction, and thus we continue to rely on speech communication throughout our lives. Because people are so comfortable with speech, there is a large desire for being able to interact with machines by speech communication (Samsudin et al., 2003).

Over the last 50 years, speech technology has advanced considerably. There are systems that interact through speech; systems that listen to what was said, and/or speak in natural language, using spoken language generation (Holmes and Holmes, 2003).

This development has also brought about an increased growth of speech recognition applications over the past years, ranging from high priced, limited dictation systems to affordable products capable of understanding

natural speech operating both at home and in professional environments [Spaans, 2004].

As a result of this advancement, speech technology is changing the way information is accessed, tasks are accomplished, and business is done. An important factor in this growth is the mobile telecommunications industry, which provides demand for speech recognition systems and also drives the development of signal processing technology (Spaans, 2004).

The ability to automatically transcribe speech signals together with its wide range of potential applications has motivated research in automatic speech recognition since the 1950s (Juang and Rabiner, 2004). Great progress has been made so far, especially since the 1970s, using a series of engineered approaches such as template matching, knowledge engineering and statistical modeling (Zhang, 2001). However, computers are still nowhere near the level of human performance at speech recognition and it appears that further significant advances will require some new insights (Zhang, 2001).

The machine, however, dictates the user interaction and requires the user to adapt to its unnatural and often complex ways. Spoken language technology will enable access to machines to become faster and easier. It is the most natural interface method possible for humans (Spaans, 2004).

1.2. Overview of Speech Recognition System

The lucrative features of the speech and ever flourishing technical ability in this field have fascinated engineers and speech scientists to think over the

utilization of speech in many new spheres of the daily life. Consequently, the topic “automatic speech recognition” (ASR) has evolved which aims to construct a machine that can emulate the human performance through recognizing and synthesizing the human speech (Podder, 1997).

Automatic speech recognition is a process by which a machine identifies speech. The machine takes a human utterance as an input and returns a string of words, phrases or continuous speech (discourse) in the form of text as output. The conventional method of speech recognition insists in representing each word by its feature vector and pattern matching with the statistically available vectors using HMM or neural network (Radman, 1999). As Furui (2001) noted automatic speech recognition system attempts to automatically extract and determine linguistic information conveyed by a speech wave using computers or electronic circuits. The linguistic information is the phonetic information that is uttered consciously to convey some meaning.

Most literature discloses that there are at least two main areas of research in relation to man machine interaction using speech (Holmes and Holmes, 2003). These are speech synthesis and speech recognition. Speech synthesis is the process of generating spoken language succession from arbitrary text whereas speech recognition is the process of converting spoken language into computer understandable text or translate it as command.

1.2.1. Feature extraction techniques

The main objective of features extraction is to extract characteristics from the speech signal that are unique, discriminative, robust and computationally efficient to each word which are then used to differentiate between different words(Ursin,2002). According to Martens (2000), there are three speech features extraction techniques. These are Linear Predictive Coding, Perceptual Linear Coding and Mel-Frequency Cepstral Coefficient.

1.2.1.1. Linear Predictive Coding (LPC)

It is one of the oldest and most simplified techniques for speech production, which works at low bit-rate inspired by observations of the basic properties of speech signals and represents an attempt to mimic the human speech production mechanism (Zhang, 2001).

The basic idea behind LPC analysis is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples. This analysis can be done through minimizing the sum of squared differences between the actual speech samples and linear predicted values. As a result a unique set of parameters or predictor coefficients can be determined.

1.2.1.2. Perceptual Linear Prediction (PLP)

Another popular feature set is the set of perceptual linear prediction (PLP) coefficients used for deriving a more auditory-like spectrum based on linear predictive analysis(LPA) of speech, which is achieved by making some engineering approximations of the psychophysical attributes of the human

hearing process(Ursin, 2002). PLP introduces an estimation of the auditory properties of human ear, and takes features of the human ear into account more specifically than many other acoustical front end techniques (Harmensky, 1990).

1.2.1.3. Mel-Frequency Cepstral Coefficient (MFCC)

According to Zoric (2005), MFCC is an audio feature extraction technique which extracts parameters from the speech similar to ones that are used by humans for hearing speech, while at the same time, deemphasizes all other information. MFCC is the most commonly used techniques in automatic speech recognition and speaker verification applications as it takes the characteristics of human auditory system into account (Zhang, 2001).

1.2.2. Types of Automatic Speech Recognition Systems

Researchers like Cook (2003) and Furui (2001) summarized major classifications in speech recognition systems since 1980s depending on the flow of speech, size of vocabulary and number of speakers. These are:

- Discrete vs. Continuous
- Speaker dependent vs. Speaker Independent
- Large Vocabulary vs. Small vocabulary speech recognition systems

1.2.2.1. Discrete vs. Continuous Speech Recognition Systems

Discrete speech recognition systems are systems that require the speaker to pause briefly between words (Zue and Cole, 1995). The pause is required to help the processor identify word boundaries and to prevent crossword co-articulation from distorting the acoustic pattern of the word to be recognized (Markowitz, 1996).

On the other hand, speech is said to be continuous when it is uttered as a continuous flow of sounds with no inherent separations between them (Markowitz, 1996). Under normal circumstances, speech is made up of continuous sounds. It is the inherent intelligence and the ability of the human mind to separate speech to intelligible words that make us think speech sound is made up of distinct words.

1.2.2.2. Speaker Independent vs. Speaker Dependent Speech Recognition Systems

Speaker-independent systems are designed to be used by virtually any user who speaks the specific language and wants to use them with no enrollment. Therefore, the reference models for these applications must recognize large, heterogeneous populations of speakers of the language (Markowitz, 1996).

However, given the many sources of inter-speaker variability the design of a speaker independent recognition system is by far more complex than a speaker-dependent recognition. On the other hand, a speaker dependent system is a system where the speech patterns are constructed (or adapted) to a single speaker (Philip and Young, 1987).

1.2.2.3. Large Vocabulary vs. Small Vocabulary Recognition Systems

Studies in speech recognition can be conducted on Small, Medium or Large Vocabularies. Although there is a variations in the division of vocabulary sizes, speech recognition systems based on large vocabulary mean that the systems have a vocabulary of roughly 1000 words or more where as small sized vocabulary recognition systems have less than 100 words or

utterances. Ultimately, recognizers with a vocabulary size ranging from 100 to 999 words are considered as a medium sized speech recognizers (Solomon et al., 2008). More technical and visual representation of the components of automatic speech recognition systems are shown in figure 1.1 below.

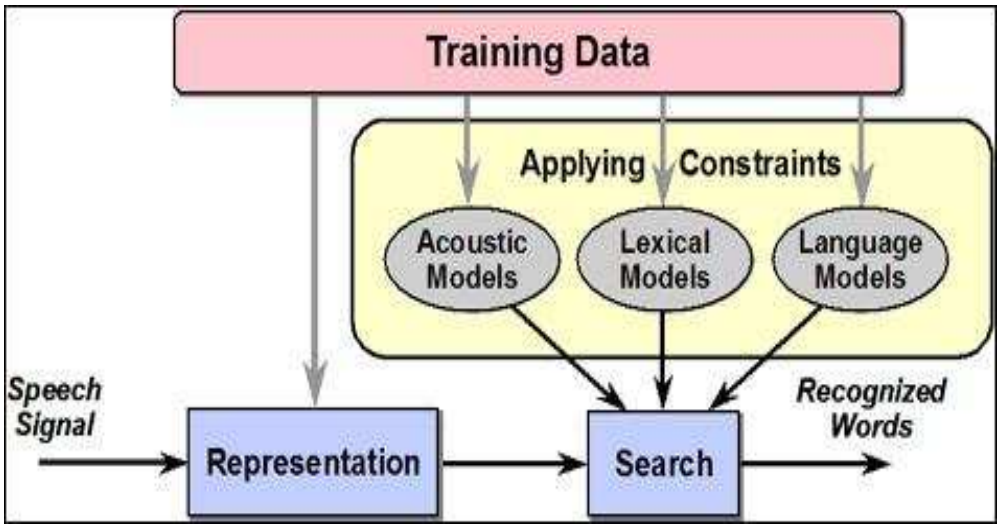

Figure 1.1: Components of speech recognition systems (adopted from Zue et al., 2001)

1.2.3. Approaches in Developing Speech Recognition

Although different scholars may use different categorization for developing speech recognition systems, there are four basic approaches to speech recognition (Rabiner and Juang, 1993). These are template matching, acoustic-phonetic, statistical/stochastic approaches, and Artificial intelligence approach.

1.2.3.1. Template Matching Approach

In template matching approach, unknown speech is compared against a set of pre-recorded words (templates) in order to find the best match (Rabiner, 1989). This has the advantage of using perfectly accurate word models. But,

it also has the disadvantage that pre-recorded templates are fixed, so variations in speech can only be modeled by using many templates per word, which eventually becomes impractical for developing usable system.

1.2.3.2. Acoustic-Phonetic approaches

This approach is the oldest, most straightforward and thoroughly researched method, whose principles are still used in artificial intelligence (AI) based recognizers. The approach assumes that the rules governing the phoneme variability are relatively simple and easily learnable. However, this approach has limitations like absence of well-defined automatic procedure and standard linguistic way of labeling the training speech. Due to these, the acoustic-phonetic based speech recognition is no longer considered as the most interesting approach (Reiderer, 1999).

1.2.3.3. Statistical based approaches

Statistical based approaches are those where variations in speech are modeled statistically, using automatic, statistical learning procedure, typically the Hidden Markov Models, or HMM. This approach represents the current state of the art in speech recognition. The main disadvantage of statistical models is that they must take priori modeling assumptions which are liable to be inaccurate, hampering the system performance. In recent years, a new approach to the challenging problem of conversational speech recognition has emerged, holding a promise to overcome some fundamental limitations of the conventional Hidden Markov Model approach Jackson (2005).

1.2.3.4. The artificial intelligence approach

This approach attempts to mechanize the recognition procedure according to the way a person applies its intelligence in visualizing, analyzing, and finally making a decision on the measured acoustic features. Expert systems are used widely in this approach (Jackson 2005).

1.3. Statement of the Problem and Justification

As the use of information technology is becoming inevitable, man-machine interaction is becoming unavoidably important. Some users can interact with the computer using the traditional methods of a keyboard and mouse as the main input devices and the monitor as the main output device. Some potential users such as physically handicapped and illiterates can not be able to interact with machines using a mouse and keyboard devices Radman (1993), hence the need for special devices. Building speech recognition systems help these users who in one way or the other are not able to use the traditional input and output devices.

Rabiner and Juang(1993), stated that speech recognition is a rich field for both practical and intellectual reasons. Practically, it could be able to solve problems of physically disabled people and improving the human computer interaction interface mechanism. Intellectually, it imposes a challenge and it is currently an active and attractive research area.

Having aforementioned general importance of speech recognition systems, Zegaye (2003) has stated that the systems that have been developed so far are meant to serve a specific language and totally limited to some 'techno-

rich' countries of the world. As Jurafsky and Martin(2000) indicated, an important distinction between speech recognition systems and other data processing systems is their use of knowledge of language (a mere acceptance of one's language recognizer for the other will not work). This language dependency of speech recognition systems ultimately needs critical looking and researching in a particular language.

When this comes to particular cases in countries like Ethiopia, many local languages which are used by a significant number of speakers were not developed through research for many reasons. Sidaama is one of such languages, which belongs to a Cushitic languages family of Ethiopia and spoken by more than 3 million people in the southern part of Ethiopia (Census, 2008).The estimates of the population vary greatly because it is not agreed where the boundaries of the language are.

Although Sidaama is spoken in relatively large area than other related Highland East Cushitic (HEC) languages such as Kenbata , Hadiya, Gedio and Burgi, there are no significant differences among different areas of the speakers. This makes the language to be very similar in dialectics. Therefore, developing speech recognizer for the language would solve many problems for the people related/attributed to education, communication, storage and retrieval of knowledge that is not available in written form.

On the other hand, the Sidaama language is being used as a language of primary and secondary schools in the region. It also serves as the official, working and administrative language of the Sidaama zone. Hence, developing

a prototype speech recognizer for the language provides a number of benefits. It can be used as an assistive tool for handicapped people who can not use keyboard and pointing devices. It can also be an ideal tool for visually impaired people if it is combined with speech synthesis.

So far, attempts have been made for developing speech recognition system for Amharic language, Afaan Oromo, and Tigregna languages. Some of research works in Amharic speech recognizer include: Solomon (2001), Kinfe (2002), Zegaye (2003), and Martha (2003). Furthermore, recently Ashenafi (2009) and Hafte (2009) tried to develop speech recognizer system for Afan Oromo and Tigregna language respectively. However, as to the best knowledge of the researcher, there are no known attempts in the area of speech recognition for Sidaama language.

Therefore, this study aims at investigating and proving out the possibility of developing a prototype speaker dependent Sidaama speech recognition system using Hidden Markov Model (HMM).

1.4. Objectives of the Study

1.4.1. General Objective

The general objective of this research was to explore the possibility of developing a speaker independent Automatic Speech Recognizer for Sidaama language that enables Sidaama speech utterances to be transcribed into a computer understandable text.

1.4.2. Specific Objectives

To accomplish the above general objective, the following specific objectives were targeted

- reviewing related literature on phonology of Sidaama language and identify list of Sidaama phonemes.
- conducting a comprehensive literature review on Automatic Speech Recognition in general and on how to develop a recognizer for Sidaama language in particular
- designing a prototype Sidaama speech recognition system with modules such as language model, acoustic model, pronunciation dictionary and decoders.
- selecting algorithms and approaches suitable to design speech recognition for Sidaama language
- evaluating the performance of the speech recognition prototype and report the findings of this research work.
- draw useful concluding remarks and suggest recommendations for further study in the area.

1.5. Research Methodology

To achieve the objectives of this research, the following approaches have been followed.

1.5.1. Review of Related Literature

Literatures such as books, journal articles, conference papers and other relevant documents have been reviewed to investigate the underlying principles/theories of the various approaches, techniques and tools that were employed in the research.

In addition, comprehensive literatures on the Sidaama language, especially those dealing with the phonetic features and articulation of the sounds of

language are examined. As a result, 58 phonesets of the language are identified and are used for building the acoustic model of the language. Moreover, to learn what others have done in the area and to better understand the problem, a comprehensive investigation of the available empirical literatures done so far on automatic speech recognition is also critically consulted.

1.5. 2. Data Collection and Preparation

The first stage of any speech recognizer development project is data preparation. Speech data is needed both for training and testing of the prototype speech recognizer model built in this research. To this end, since there is no prerecorded speech corpus available for this language, the researcher recorded speech datasets. In order to represent the phonetic features of the language and to make the words selected as varied as possible a domain linguistic expert is consulted and as a result 450 representative words are selected.

The speech data for the research is recorded from a single speaker. 2/3 of the recorded speech data is used to train and construct the speech models and the remaining 1/3 of the recorded speech data is used for testing the accuracy of the recognizer models. A total of 450 words selected by an expert were uttered by the selected speaker while it was recorded with a sampling frequency of 1600Hz and mono mode headset microphone. In addition, in order to reduce external disturbances and to maintain the quality of speech data used for the study, the recording was done in an office room.

The training data is used during the development of the system. Test data provides the reference transcriptions against which the recognizer's performance can be measured and a task grammar is used as a random generator to create them. In the case of the training data, the prompt scripts is used in conjunction with a pronunciation dictionary to provide the initial phone level transcriptions needed to start the HMM training process.

It follows from above that before the data can be recorded, a phone set must be defined, a dictionary must be constructed to cover both training and testing, and a task grammar has to be defined.

1.5.2. Modeling Technique used for the Study

The modeling technique used in this research is the Hidden Markov model (HMM).The reason for the selection of the Hidden Markov Model from other modeling techniques is that it has been successfully used in speech recognition for many years (Plannerer,2005). Furthermore, previous researchers Solomon (2001), Kinfe (2002), Martha (2003) employed this model and achieved promising results.

Lee (1989) has also stated that the Hidden Markov Model is a powerful and predominant technique capable of robust modeling of speech and a parametric model that is particularly suitable for describing speech events.

1.5.4. Development Tools

Selection of appropriate development tool for the research has a role to play for the success of the study.

A speech acquisition and labeling tool called Praat[1] is used to undertake the data acquisition process. Praat is an open-source software tool for speech signal edition and labeling. It is chosen among other speech signal acquisition tools because it is specifically designed to handle speech signal analysis and related scientific experiments with greater ease of use.

The experimentation conducted in this research work has utilized state of the art open source speech recognition and processing tool called Sphinx version4, which is entirely built in Java programming language. The tool provides a lot of important features such as code reusability and ease of experimentation as compared to other recognition tools. This tool is the result of an open source project by Carnegie Mellon University (CMU), SUN Microsystems Inc. and Mitsubishi Electric Research Laboratory (Walker et al., 2004). It is based on a highly modularized and flexible architecture which supports various techniques and approaches in speech recognition and could easily be extended or modified for further research.

Once the input corpus is collected using Praat, building of the acoustic model from the data was carried out by one of the tools available in the Sphinx system called SphinxTrain. This particular tool takes the utterances recorded using Pratt to result in an acoustic model that is convenient for a font-end speech recognizer tool called Sphinx-4 which was selected for the study.

[1] Praat is a Speech recording and labelling tool, It can be downloaded from: http://www.brothersoft.com/praat-171746.html

1.5.5. Testing and Evaluation

Once the HMM-based recognizer is built, it is necessary to evaluate its performance. This is usually done by using it to transcribe some pre-recorded test words and match the recognizer output with the correct reference transcriptions(the ground truth).

The most important and common testing parameter used to evaluate speech recognition system is the word level accuracy of recognition (Zegaye, 2003). This research has also used the word error rate (WER) in order to measure how well the recognizers classified the input acoustic information to the corresponding text words.

1.6. Application of the Study

Conducting a research on Automatic speech recognition system for Sidaama is worth undertaking for various reasons apart from the academic challenges that it poses to qualify as thesis.

This work paves the way in identifying some of the possible ways of developing speaker dependent speech recognizer for Sidaama language. This work in turn could be further extended with future researches to improve the existing man-machine interactions particularly pertaining to the speakers of the language.

The findings of this work can also be applicable for developing a full fledged Sidaama speech recognizer. It can also be used for the general applications

of an automatic speech recognizer such as dictation, command and control, telephony and use for disabled people.

In general, apart from being a natural way of interfacing with machines like PCs, Telephones, Television etc, Automatic speech Recognition provides the following benefits:

- helps to speed up computer data entry (much faster than using the ubiquitous, keyboard even for the fastest possible typists);
- avoids pains related to typing (like repetitive stress injury); using ones voice, the hands and the eyes are free and movement is unconstrained. The same can't be said of systems that rely on keyboards, mice, or touch screens for input.
- the technology will find applicability in systems such as banking, telecommunications, transport, Internet portals, accessing PC, emailing, administrative and public services, cultural centers and many others

Since speech technology is the technology of today and tomorrow the result of this research can be taken as a good start to help the speakers of the language to take many advantages of ICT.

1.7. Scope and limitations of the Study

The scope of this research is limited to developing speaker dependent speech recognition of Sidaama language. This research also aims to build a medium vocabulary size based prototype speaker dependent recognizer. The

developed prototype systems integrated different components such as acoustic models, language model, and decoder and recognition units.

Most studies showed that, it is more convenient to deal with isolated word recognizers than continuous speech recognizers, as the problems of coarticulation and identifying word boundaries could be avoided. Hence, in this particular research, both context dependent and context independent isolated word recognizers were considered. This research also employed triphones as recognition units in building context dependent model as well as monophones as recognition units in developing context independent model.

Lastly, comparative analysis of the two contending recognition units was made based on the performance of the respective recognizers, and the recognition unit which proved to be strong contender is selected. Finally, conclusions were made based on analysis of the results obtained from the prototype models.

The main limitation of the study was absence of ready made Sidaama speech corpus to be used for this research. Hence, selecting phonetically balanced words as well getting an individual who is willing to participate in the research for recording his/her utterances was challenging.

Human and environmental factors had also their own impact on the results of the experiments because some of the speakers were stressed and were not consistent in their utterances and the recording environment selected is not

primarily designed for the purpose of recording speech, hence it had its won negative impact on the quality of the recorded data.

Shortage of time was also another limitation that faced this work. Had there been more time, more performance optimization tasks could have been done as well as java based interface for the recognizer models could have been developed.

1.8. Organization of the Thesis

This thesis is organized into six chapters. The first chapter presents an introduction to the subject matter under study, description of the statement of the problem along with justifications and objectives of the research. It also presents an overview on speech recognition, methodologies adopted on the research, applications of speech recognition systems and beneficiaries of the new system.

Chapter two provides literature review on human speech production system, overview of speech recognition and also different techniques of speech recognition are discussed in detail. The Sidaama phonetics are discussed in chapter three that describe the characteristics and way of creation of the Sidaama phonemes both consonants and vowels followed by its syllable structure.

Chapter four explains the Hidden Markov Model (HMM), elements of HMM, assumptions in HMM, Problems of HMM and types of algorithms used in HMM. The basic components and features of the Sphinx speech recognition system are also discussed in this chapter.

Chapter five discusses experimentation and testing of the Sidaama prototype speech recognizer in detail.

Lastly, chapter six provides concluding remarks and recommendations for future research in the area.

CHAPTER TWO

THE SIDAAMA LANGUAGE

Sidaama (*sidaam-u afo* in Sidaama) is one of the East Highland Cushitic language families. It is a branch of the Afro-Asiatic language phylum spoken in the Sidaama Zone and some parts of the Southern Nations, Nationalities, and People's Region of Ethiopia. (Ambessa, 2000; Kawachi, 2007). The term 'Sidaama' is used as the name by which the people refer to themselves, their region and their language (Kawachi, 2007).

Although Sidaama is spoken in a relatively large area than other related Highland East Cushitic (HEC) languages such as Haddiya, Gedeo, and Burji, there are no significant differences among speakers of different areas of Sidaama (Ambessa, 2000; Kawachi, 2007). Confirming this fact Hudson (2005), described Sidaama language as wide spread language of the HEC group with little difference in dialects.

Ambessa (2000) also stated that, out of the Cushitic languages spoken in the Ethiopia, Afaan Oromo, Somali, Sidaama, Hadiya, and Afar-Saho are the languages with the highest number of speakers.

2.1. The Sidaama Writing System

Until the mid 1970's Sidaama was a vernacular language (Kawachi, 2007). It was not confined and as such was not used for teaching or other purposes. However, the Military Regime of Ethiopia (MRE) permitted Sidaama language to be used for the ongoing literacy campaign. During this period Sidaama was written using the Ethiopian script. However, since localization was done

inaccurately, reading and understanding the Sidaama documents and books was very difficult. This emanated from the discrepancy in the number of vowels in Amharic and Sidaama and inappropriate decision regarding the representation of vowels (Ambessa, 2000). However, the selection of Sidaama as a language of literacy enabled the publication of several educational primers.

Another fundamental change with regards to the status of Sidaama language took place in 1992. The Transitional Government of Ethiopia (TGE) at the time declared that each and every ethnic group can use its own language for primary education and also for official and administrative purposes. In addition permission was given to use any type of script in order to write the vernacular languages. Thus, beginning from 1992, Sidaama is being used as a language of primary education, and for administrative and judicial matters. In addition the Latin script was adopted as the writing system (Kawachi, 2007). The following list shows the Latin Fidela used in Sidaama language with their respective pronunciations (Ambessa,2000).

A a	B b	Cc	CH ch	D d	DH dh	E e	F f	G g	H h	I i	J j	K k	L l	M m	N n	NY ny	O o	PH ph	Q q	R r	S s	SH sh	T t	U u	W w	X x
[ɑ]	[b]	[tʃ']	[tʃ]	[d]	[ɗ]	[e]	[f]	[g]	[h]	[i]	[ʤ]	[k]	[l]	[m]	[n]	[ɲ]	[o]	[p']	[k']	[r]	[s]	[ʃ]	[t]	[u]	[w]	[t']

Y y	Z z
[j]	[z]

Table 2.1: Sidaama Fidella and the IPA symbols for Sidaama (Kawachi, 2007)

2.2. The Phonology of Sidaama

Phonology is the study of the distribution and patterning of speech sounds in a language and of the tacit rules governing pronunciation (Roach, 2002). Similarly, Ambessa (2000) defined phonology as a branch of linguistics that deals with the system of speech sounds in a language. In phonology, phoneme is the fundamental unit that describes how speech conveys linguistic meaning. The phoneme represents a class of sounds that convey the same meaning. The meaning of a word is dependant on the phoneme that it contains (Olsen, 2003).

Phones are divided into two main classes: consonants and vowels. Both kinds of sounds are formed by the motion of air through the mouth, throat or nose. Consonants are made by restricting or blocking the airflow in some way, and may be voiced or unvoiced. Vowels have less obstruction, are usually voiced, and are generally louder and longer-lasting than consonants (Jurafsky and Martin, 1999).

2.3. Consonant Phonemes

In principle, two sounds are considered to be separate phonemes if they bring change of meaning in a pair of words. The following are some of the minimal pairs used to identify the Sidaama consonant phonemes. Based on the contrastive minimal pairs 25 consonant phonemes are identified for Sidaama. Table 2.2 presents such phonemes of the Sidaama language (Kawachi, 2007).

Consonants						
Place of Articulation		**Manner of Articulation**				
		labial	**Alveolar**	**Palato-alveolar**	**Velar**	**Glottal**
Plosives	**Voiceless**		t	ch /ʧ/	k	? /'/
	Voiced	b	dh /d/	J	g	
	Ejective	P' /ph/	X /t'/	c /ʧ'/	Q /k'/	
	Implosive		d			
Fricatives	**Voiceless**		s	sh / ʃ /		h
	Voiced		z			
Affricative	**Voiceless**	f				
	Voiced					
	Ejective					
Nasal		m	n	ny /ɲ/		
Lateral			l			
Flap			r			
Semivowel		w		Y [j]		

Table 2.2: Sidaama consonant phonesets (adopted from Abebe, 1985

The language has a series of ejectives (/ P'/, /t'/, /k'/, /c/) like other Eastern Cushitic Languages (Hudson, 2005). Although the favored place of articulation for ejectives is the back of the mouth in many languages including a number of Cushitic languages, Sidaama has /p'/ or PH . On the other hand, unlike the ejectives, which are single segments, the glottalized sonorants (/'l/,/'m/,/'n/,/'r/,/'y/ are clusters made up of two phoneme segments(Abebe,1985). The glottal stop is phonemic in this language.

Sidaama lacks /p/ and /v/. According to Kawachi (2007),these two consonants are rarely used in Ethiopian languages, and in loan words,/p/ and /v/ are usually replaced by /f/ and /b/, respectively. For example, the

word ' politics' in English ('ፖለቲካ' in modern Amharic) is pronounced as 'foletika' in Sidaama native speakers. Similarly, the word '*university*' in English ('ዩኒቨርሲቲ' in modern Amharic) is pronounced as *'yunibersite*' in native Sidaama speakers which is pronounced in (Amharic as 'ዩኒበርሲቱ').

Consonants can occur both as single and also as geminates. Gemination is phonemic in Sidaama (Ambessa,2000). This indicates that all consonant phonemes of Sidaama have a geminate counter parts as the two examples below demonstrate.

/gato/ meaning 'forgiveness' /gatto/ meaning 'poverty'

/ada/ meaning 'father's sister' /adda/ meaning 'truth'

The plato-alveolar stop /č/occurs only as a geminate as in /guluччo/ meaning 'knee' or as a second member in hetreosyllebic sequence as in /danča/ meaning 'good'.

In most instances, the pre-glottalized flap /'r/ is an inetrvocalic realization of the phoneme DH /ɗ/. Hudson (2005) confirmed this fact by saying'' Sidaama /r/ is an alternate of intervocalic DH''. The following example illustrates this distribution.

/hadh-a/ and /*ha'r-a*/ which means *' to go '* in English

However, there are two words where the segment / 'r/ is found lexeme-internally and not a phonemic realization of /DH/. These words are /go'ra/ 'raspberry' and /ba'raa'ra/ 'joy'.

The alveolar fricative /z/ is found only in Amharic loan words such as /zufanne/ meaning 'throne' and /muuze/ meaning 'banana'.

2.3.1. Vowel Phonemes

One of the basic differences between voiceless consonants and vowels is that in the case of vowels, the air generated by lung will vibrate the vocal cord since the vocal cord is slightly closed when vowels are generated. Teferra(2000) defines vowels as speech sounds produced without an obstruction of the airflows in the oral cavity.

Sidaama has five short distinct vowel phonemes: /i/, /e/, /a/, /o/, and /u/ as well as five long contrastive long vowel phonemes: /ii/, /ee/, /aa/, /oo/, and /uu/. Vowel length is phonemic in Sidaama (Abebe,1985). Examples of such minimal pairs that are in contrast in vowel length are shown in table 2.3.

Pair of Vowels	**Sidaama words**	**English Equivalent**
(i, ii)	sinna	Branches
	siinna	coffee cups
(e, ee)	tenne	at that time
	teenne	flies
(a, aa)	j awa	great, old
	j aawa	to be come thin
(o, oo)	k' ola	to reply
	k' oola	wing
(u, uu)	kula	to tell
	kuula	blackish blue

Table 2.3: Sidaama Vowel Minimal pairs

Similarly, the tongue plays an important role in generating different vowels. The contribution of tongue can be seen in two different aspects. The first one is the way the tongue is positioned in terms of height and the second one is the movement of tongue to the front and back of the mouth. In addition to tongue, the shape of lips has also a contribution in changing the vowels sound when it varies from rounded to non-rounded (flat). Moreover, vowels differ from consonants in that there is no noticeable obstruction in the vocal tract during their production. Air escapes in a relative unrestricted way through the mouth and/or nose (Olsen, 2003).

The vowels in Sidaama language can be divided into different categories depending on how they are formulated such as front, central or back position of the tongue, wideness or roundness of the constriction position,

and place of the tongue (high, mid or low) (Rodman,1999). Sidaama vowels and their categorization are presented in table 2.4 below.

	Front	**Central**	**Back**
High(short and long)	i , ii	-	u, uu
Mid(short and long)	e, ee	-	o, oo
Low(short and long)	-	a, aa	-

Table 2.4: Sidaama vowel phonemes and their categorization

All Sidaama words end in vowels (Kawachi, 2007). Generally open class words end in /e/, /a/, or /o/ in their citation forms and can end in /u/ or /i/ only when followed by a suffix that consists only of or ends in one of

these vowels (Kawachi, 2007). The vowel phoneme /i/ is usually realised as a high front vowel in Sidaama language (Abebe, 1985)

2.4. Syllable Structure

A syllable is a unit of phonological organization composed of one or more segments and minimally containing a nucleus, usually a vowel. The Sidaama language has both closed and open syllables. Closed syllables in the language occur ward internally as in words such as /wo-ba/ meaning 'brave' and /kub-ba/ meaning 'to jump'. However, open syllables tend to dominant in Sidaama since every inflected word ends in vowel (Kawachi, 2007).

Possible syllable structures of Sidaama can be formulated as V, VV, VC, VCC, CV, CVV, CVC, CVVC, and CVCC in which C stands for a consonant and V for a vowel. The following are example words which contain the syllable structures mentioned above respectively.

Syllable structure	**Sidaama Word**	**English Equivalent**
V	a.do	'milk'
VV	aa.na	'top, to become equal'
VC	im.ba.ra	'the day after tomorrow'
VVC	iib.ba.do	'hot'
CV	mi.ne	'house'
CVV	ka'a	'rise or get up'
CVC	han.do	'ox'
CVVC	haan.ja	'green'

Table 2.5: Example words with syllable structure

2.5. Consonant Sequences

The maximum number of consonants that can occur successively in Sidaama language is two and consonant clusters only occur intervocalically across syllables (Kawachi,2007). All Sidaama consonant sequences are hetro-syllabic. Kawachi(2007) described three types of such sequences with their respective illustrative examples.

- *Sonorant-obstruents* **:**

the nasals and liquids can each be followed by immediately obstruents. The following demonstration describes such pattern in Sidaama language.

/m/- obstruent:

/mb/ ambooma 'hayena' /mf/ gimboola 'bamboo basket'

/n/- obstruent:

/nt/	onte	'five'	sante	'coin'
/nd/	hando	'ox'	danda	'be able to do'
/nk/	hinko	'tooth'	dunka	'slow'
/ng/	anga	'place'	gongoma	' to roll'

/r/-obstruent:

/rb/	warba	'smart'	darbata	'heavy cloth'
/rt/	t'eerto	'far'	amaartichcha	'male Amhara person'
/rg/	darga	' place'	dargara	'to drive away'

/l/-obstruent:

/lb/	alba	'face'	bulbula	' to mix with water'

/lt/	salto	' liver'	t'lte	' waasa container'
/ld/	belda	'rich'	k'alda'	' storage'

- *Sonorant-sonorants:*

There are small numbers of such words in Sidaama and some of them are loan words. Examples of such words are shown below and the dots indicate syllable boundary.

/baay.ra/ 'elder' /kaay.ni/ 'but' /oy.loo/ 'forty'

- *Glottal-sonorant:* examples below show such words in Sidaama.

/kaa'.la/ 'to help', /ma'na/ 'bed' and /go'ra/ 'raspberry'

2.6. Supra-segmental

Tone is phonemic in Sidaama (Kawachi, 2007). Words that are spelled similar may pronounced differently and have different meaning. Examples are mentioned below.

/zare/ means lizard and relative.

/marra/ means arrange orderly and stir it up.

/maata/ means grass and right.

Hence, form the above discussion of the Sidaama language phonetic features and structure it can be concluded that, unless understanding the phonology of Sidaama in detail, developing the prototype Speech recognizer for this language is difficult.

CHAPTER THREE

LITERATURE REVIEW

Humans have the ability to communicate with other humans in various ways, which includes, body gestures, the printed text, pictures, drawings, and voice. However, voice communication is used widely in our daily activities. Since speech has been demonstrated to be the most natural, effective and efficient way for humans to express ideas and requests, it does not come as a surprise that a desire exists to communicate with machines by voice (Shaefer, 1995).

Hence, the goal of speech recognition technology, in a broad sense, is to create machines that can receive spoken information and act appropriately upon that information. From this perspective, the study of speech recognition is part of a quest for artificially intelligent machine that can understand and act upon spoken information (Shaefer, 1995).

3.1 History of Speech Recognition

Designing a machine that mimics human behavior, particularly the capability of speaking naturally and responding properly to spoken language, has intrigued engineers and scientists for centuries Juang and Rabiner(2004).

Since the 1930s, when Homer Dudley of Bell Laboratories proposed a system model for speech analysis and synthesis, the problem of automatic speech recognition has been approached progressively, from a simple machine that

responds to a small set of sounds to a sophisticated system that responds to fluently spoken natural language and takes into account the varying statistics of the language in which the speech is produced (Furui, 2005).

Early attempts to design systems for automatic speech recognition were mostly guided by the theory of acoustic-phonetics, which describes the phonetic elements of speech which are the basic sounds of the language, and varies researchers' tried to explain how they are acoustically realized in spoken utterance. These elements include phonemes and the corresponding place and manner of articulation used to produce the sounds in various phonetic contexts (Juang et al., 2004).

The first machine to recognize human speech was a toy - a celluloid dog, the "Radio Rex". The simple electromechanical toy from 1920 was capable of jumping, when its name was spoken. Since then, speech recognition has been a topic of great interest and a lot of attempts have been made in this area by researchers around the world (Juang and Rabiner,2004).

The vast potential of speech recognition and its advantages was noticed by Unite States Department of Defense as early as the late 1940's. During this period the United States Department of Defense was sponsoring development of an automatic language translator capable of automatic translation of intercepted Russian messages. However, this project was a large failure because creating a program that could recognize speech was too difficult due to low speed computers that were available at the time (Juang and Rabiner, 2004).

During the 1950's, the first system capable of recognizing digits spoken over the telephone was developed by Bell Laboratories. The built system was an isolated digit recognizer for a single speaker using a formant frequencies estimated during vowel regions of each digits(Lee and Juang, 1996).

In the 1960's, several Japanese laboratories demonstrated their capability of building special purpose hardware to perform a speech recognition task. Most notable were a hardware vowel recognizer built at the Radio Research Lab in Tokyo and the phoneme recognizer built at Kyoto University (Juang and Rabiner, 2004 and Anusuya and Katti,2009).

Another hardware effort in Japan was the work of Sakai and Doshita involving the first use of a speech segmenter for analysis and recognition of speech in different portions of the input utterance. A hardware speech segmenter was used along with a zero crossing analysis of different regions of the spoken input to provide the recognition output(Anusuya and katti,2009).In contrast, an isolated digit recognizer implicitly assumed that the unknown utterance contained a complete digit (and no other speech sounds or words) and thus did not need an explicit "segmenter." Kyoto University's work could be considered a precursor to a continuous speech recognition system.

Similar early recognition system a phoneme recognizer with a capability of recognizing 4 vowels and 9 consonants was built in England. Later on by incorporating statistical information about allowable phoneme sequences in English, the overall phoneme recognition accuracy for words consisting of

two or more phonemes is increased. This work marked the first use of statistical syntax (at the phoneme level) in automatic speech recognition (Juang and Rubiner, 2004).

Since the late 1970's, mainly due to the publication by Sakoe and Chiba, dynamic programming, in numerous variant forms including the Viterbi algorithm has become an indispensable technique in automatic speech recognition (Lee and Juang, 1996).

In the 1970s speech recognition research achieved a number of significant milestones. First the area of isolated word or discrete utterance recognition became a viable and usable technology based on fundamental studies in Russia, Japan, and the United States (Anusuya and Katti, 2009).

The Russian studies helped the advance use of pattern recognition ideas in speech recognition; the Japanese research showed how dynamic programming methods could be successfully applied; and Itakura's research showed how the ideas of linear predictive coding (LPC) could be extended to speech recognition systems through the use of an appropriate distance measure based on LPC spectral parameters (Juang and Rabiner, 2004).

A Speech research in the 1980s was characterized by a shift in technology from template based approaches to statistical modeling methods especially the hidden Markov model (Juang et al.,2004). During this period, researchers started to tackle large vocabulary speaker-independent, continuous speech recognition problems. Speaker-independence has been possible because of HMM’s ability of capturing and modeling speech

variability. Artificial neural networks (ANN) were also reintroduced in the late 1980's. Similarly, recording of large speech databases such as TIMIT, which contributed to the advances made in ASR, was done in this period (Solomon et al., 2008).

Based on major advances in statistical modeling of speech in the 1980s, automatic speech recognition systems today find widespread application in tasks that require a human-machine interface, such as automatic call processing in the telephone network and query based information systems that do things like, provide updated travel information, stock price quotations, weather reports and so on (Solomon et al.,2008).

In the 1990s it was possible to build large vocabulary systems with unconstrained language models, and constrained task syntax for continuous speech recognition (Juang and Rabiner, 2004). The success of statistical methods revived the interest from Department of Advance Research Projects Agency (DARPA) leading to several speech recognition systems such as the Sphinx system, the BYBLOS system and the DECIPHER system from CMU, BBN and SRI respectively. As a result, researchers developed an increasing interest for speech processing under noisy or adverse conditions, spontaneous speech recognition and dialog management (Lee and Juang, 1996).

In the first half of 2000s, the attention of researchers has been attracted towards spontaneous, robust, multimodal speech recognition. Although accuracy of more than 95% can be achieved using read speech, the accuracy

drastically decreases for spontaneous speech. On the other hand, speech recognition systems applied for real applications should have high accuracy for spontaneous speech and should also be robust (Furui, 2005).

Human beings use multimodal communication when they speak with each other especially in noisy environments. Speech recognition systems developed using multimodal audio and visual information showed better recognition performance than systems developed using only speech information. During this period, DARPA conducted The Effective Affordable Reusable Speech-to-Text (EARS) program so as to develop automatic transcription with the aim of achieving more accurate output (Furui, 2005).

As it can be observed from the above review of literatures in speech recognition researches, speech recognition technology has gone through significant progress. The use of HMM, the development of large speech corpora for system development training and testing, establishment of standards for performance evaluation, and advances in computer technology are the factors that fueled the progress.

3.2 Basic concepts of ASR

Speech recognition technology and ASR systems have been defined differently in respect to the various, yet different, applications and domains for which they are used. Researchers and scientists have also defined speech recognition technology and ASR systems according to the way they use them in their researches. However, all ASR systems aim at automatically extracting the string of spoken words from input speech

signals.

Forsberg (2003) defines speech recognition, or more commonly known as automatic speech recognition (ASR), as the process of interpreting human speech in a computer. Jurafsky and Martin (2000) also defined ASR more technically as the building of system for mapping acoustic signals to a string of words.

Speech recognition is also defined as the technology by which sounds, words or phrases spoken by humans are converted into electrical signals, and these signals are transformed into coding patterns that can be identified by a computer. Based on this identification, the computer usually takes some actions. Speech recognition also refers to the ability of a machine or computer program to receive and interpret spoken commands and act upon those commands.

Therefore, automated speech recognition, using a microphone or a telephone as an input device, converts a person's speech into digital code by comparing the electrical patterns produced by the speaker's voice with a set of pre-recorded patterns stored in the database.

3.3 Human Speech Production Mechanism

Speech is a natural form of communication for human beings, and computers with the ability to understand speech and speak with the human voice are expected to contribute to the development of more natural man-machine interfaces. Computers with this kind of ability are gradually

becoming a reality, with the evolution of speech recognition and synthesis technologies (Honda, 2003).

However, in order to give computers functions that are even closer to those of human beings, we must learn more about the mechanisms by which humans produce speech and develop speech processing technologies that make use of these functions. Understanding the mechanisms on which speech is based also helps to clarify how the brain processes information and leads to the development of more human-like speech devices through the imitation of these functions by computers (Honda, 2003).

The main components of the human speech system are the lungs, trachea (windpipe), larynx (organ of speech production), pharyngeal cavity (throat), oral or buccal cavity (mouth), and nasal cavity (nose). (Rabiner and Juang,1993) and Holmes,(2001). Figure 3.1 shows a schematic view of human speech production mechanism

According to Rabiner and Juang (1993), the lungs and the associated muscles are considered as the source of air for exciting the vocal mechanism. The muscle force pushes air out of the lungs and through the trachea. When the glottis or vocal cords are tensed, the air flow causes them to vibrate, producing the voiced speech sounds. Whereas, when the vocal cords are relaxed, in order to produce a sound, the air flow passes through a constriction in the vocal tract and thereby become turbulent, producing the unvoiced speech sounds.

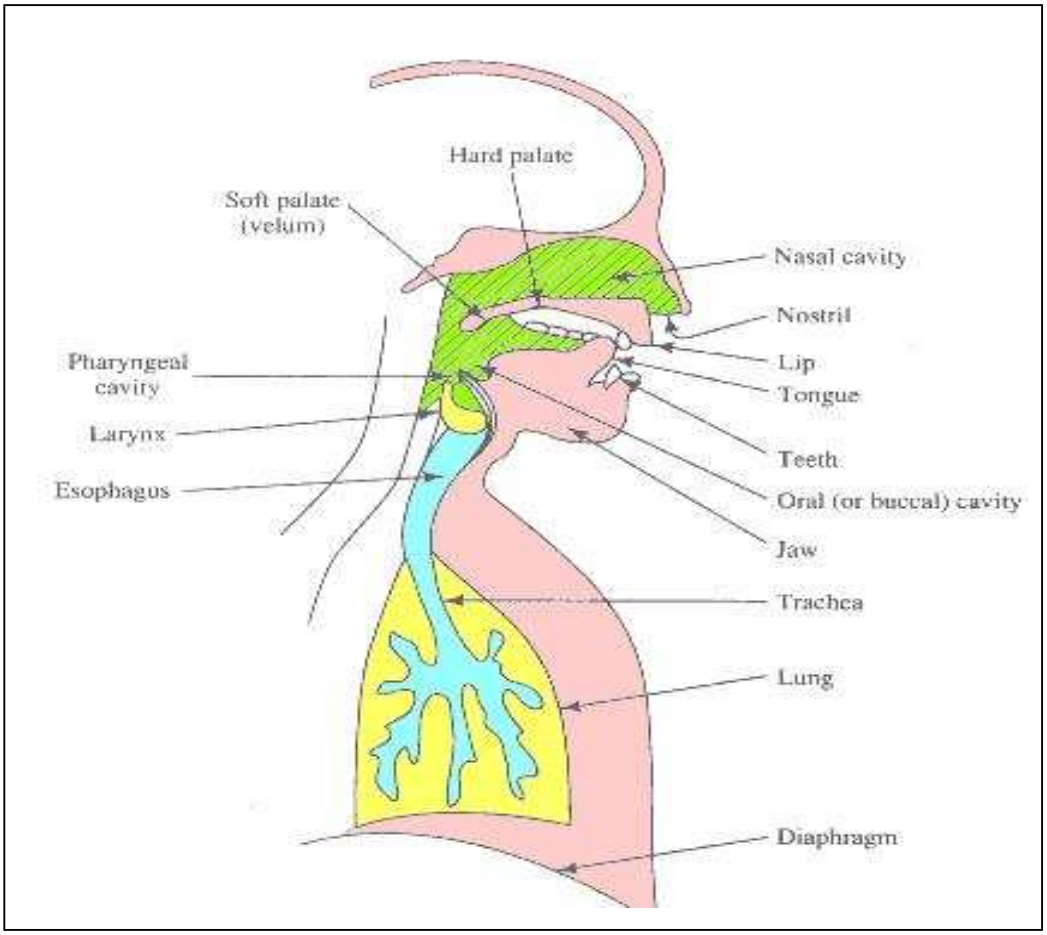

Figure 3.1: Schematic vew of human speech production mechanism
(adopted from Rabiner and Juang, 1993)

Therefore, a speech sound is produced by the vibration of the vocal cords in the case of voiced speech sounds and air turbulence in the case of unvoiced speech sounds by a cooperation of lungs, glottis (vocal cords) and articulation tract (mouth and nose cavity).

3.3.1. Speech Events, Phones and phonemes

A speech event is any kind of sound produced by the speech organs of a person. Two speech events are considered to be different if any difference can be found between them, for example, in the spectral structure or timing.

Therefore, any two speech events are most likely different - even if produced by the same person (Laver, 1990).

Speech events that are phonetically equivalent between speakers can be referred to as phones. Phonetic equivalence is a complex concept that relies on the idealizing assumption that the differences between a set of speakers' phone organs can be ignored when evaluating the speech event's phonetic quality (Laver, 1990).

Phones are regarded as realizations of the same abstract phoneme - they are its allophones or variants. It should be emphasized that a phoneme is indeed an abstract concept. It is not possible to divide a word into phones with clear, unambiguous borders because the speech signal is changing continuously. The phoneme model is simply an idealizing abstraction of the complex phenomenon called speech (Ursin, 2002).

3.3.2. Speech Recognition Units

When building a speech recognizer, an important decision one has to initially make is to choose the recognition unit. This unit can be of different lengths, and each choice has its own benefits and disadvantages. The unit should be consistent and trainable: different instances of the same unit should have similarities, and there should be enough training samples to reliably estimate the model parameters.

In order to design a speech recognizer for any language one must establish some ways of symbolizing the spoken utterances by some group of symbols that distinctly represent the sound produced. Based on the work of Lee

(1990), different recognition units with their respective explanation are described:

3.3.2.1. Word Model

Words are the actual unit we eventually want to recognize. They are thus a natural choice for the speech recognition unit. They inherently take context-dependency issues into account, since phonemes inside a word are very likely the same in each utterance. The obvious problem with word models is the large number of different words and the problem of too little training data. This makes word models less usable for large vocabulary recognition.

3.3.2.2. Phone Model

To achieve parameter sharing across different words, sub-word units like phonetic models are often used because they are both trainable and sharable. Since there are small amount of phone units, phones can be sufficiently trained with just a few thousand sentences. Hence, unlike word models, monophone models have no training problem.

Despite its many advantages, monophone assumes that a phone produced in one context is equivalent to the same phone in any context. In fact phones are not produced independently and its realization is strongly affected by its surrounding phones. Pooling all the acoustic variants of the same phone into one model produces inaccurate modeling for any particular variant.

3.3.2.3. Syllable Models

One way to share parameters and at the same time keep detailed and

consistent models is to use multi-phone units. Examples of these include syllables, semi-syllables etc. However, while the co-articulations in the middle portions of these units are well modeled, the beginning and ending portions are still susceptible to some contextual effects. The large number of multi-phone models presents another considerable problem for large-vocabulary tasks. Moreover, there is no parameter sharing between different multi-phone models.

3.3.2.4. The Diphone and Triphone Models

The biphone and triphone models are context dependent models. A left-context biphone is defined as a phone model with particular phone as its left context. Similarly, a right-context biphone is dependent on its right context. Although bi-phones are sharable across different words and take into account half of the important contextual effects, the remaining contextual effects go un-modeled. Triphone model considers both the left and right phones as its context. Each triphone model is used to represent a certain phone surrounded by two particular phones. However, they are difficult to train due to the tremendous number of possible triphones.

3.4 Components of Automatic Speech Recognition (ASR)

Visualizing the components of a given speech recognizer helps how the various subsystems that make up the recognizer are arranged and interoperate to yield an acceptable recognition output. Depending on the method followed to design a recognizer, the architecture constructed may have different components and interaction mechanisms. For instance in

simple pattern matching recognizer, building a language model may not be essential.

Accordingly, the methods and techniques used to perform a particular functionality in each and every component of the recognizers, once the over all design is set, will also greatly vary based on a particular approach followed. For instance, recognizers using statistical approaches will mostly use Hidden Markov Model (HMM), where as recognizers using Artificial Neural Networks choose artificial intelligence (AI) approaches (Comez, 2003). The principal components of HMM based speech recognizer architecture is shown in figure 3.2.

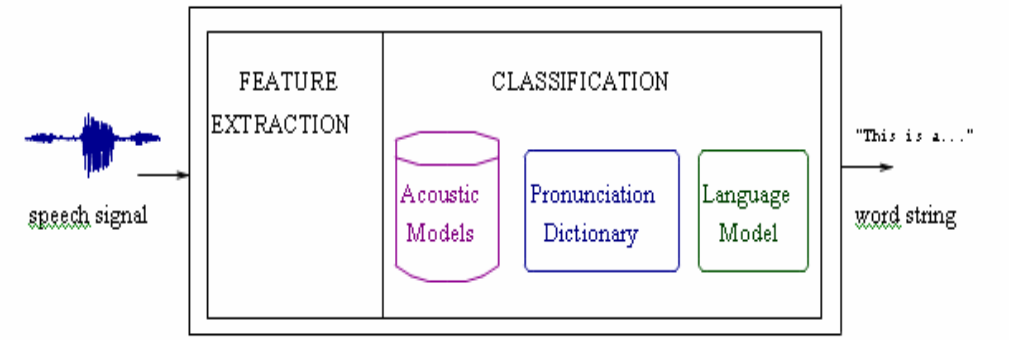

Figure 3.2: Principal components of ASR system architecture (adopted from Wiggers, 2001)

3.4.1 Feature Extraction Component

The first component of the ASR in Figure 3.2 is Feature Extraction where the sampled speech signal is parameterized. The goal is to extract a number of parameters ('features') from the signal that have a maximum information content relevant for the following classification. That means features are

extracted that are robust to acoustic variation but sensitive to linguistic content. In other words, features that are discriminant and allow separation between different linguistic units are required.

Moreover, the features should also be robust against noise and factors that are irrelevant for the recognition process. 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. Hence, the choice of suitable features varies depending on the classification techniques employed.

3.4.2 Acoustic Model Component

Acoustic models of speech recognizer are built from speech utterances and their transcriptions that are produced during the pre-processing from the speech corpus and implementing those inputs into statistical representations of the sounds that are used to produce each word. Acoustic model provides mapping between a unit of speech and an HMM that can be stored against in coming features provided by the Font End (Walker et al., 2004).

In the classification module the feature vectors are matched with reference patterns, which are called acoustic models. The reference patterns are usually Hidden Markov Models (HMMs) trained for whole words or, more often, for phones as linguistic units. HMMs cope with temporal variation, which is important since the duration of individual phones may differ between the reference speech signal and the speech signal to be recognized.

3.4.3 The Language Model Component

A language model also called grammar in command and control operation systems contains probabilities of sequence of words. Generally, a language model in continuous speech recognition systems is used to help how words could be combined to make meaningful sentences while the grammar in isolated word recognition systems is just the specification of the terms used in the system (Walker et al., 2004).

The language model module of the recognizer provides word-level language structures, which can be represented in different implementations (Walker et al., 2004). These implementations fall into two major categories:

Graph-Driven Grammars: The graph-driven grammar represents a directed word graph where each node represents a single mode and each arc represents the probability of a word transition taking place.

Stochastic N-Gram Models: the stochastic N-gram models provide probabilities for words given the observation of the previous (n-1) words.

3.4.4 Pronunciation Dictionary

A dictionary provides pronunciations for words found in the Language Model. The pronunciation break words into sequence of sub-word units found in the acoustic model (Walker et al., 2004).

3.5 Techniques of Speech Recognition

In the area of speech and language processing, a number of recognition techniques have been used and have achieved promising results. The most

common and successfully used techniques of speech recognition are described below.

3.5.1. Dynamic Time Warping (DTW)

Dynamic time warping (DTW) is one of the oldest and most important algorithms in speech recognition. DTW approach is a template matching method, where it compares the unknown pattern with its reference template to get the minimum score. The minimum score indicates that the unknown pattern is the most likely to be matched onto the particular reference template compared to other reference (Ursin, 2002).

While effective in pattern recognition, the DTW algorithm requires long processing time and large amount of storage which become the major problems for real time application as the number of speech patterns increases. As a result, it is only useful in small vocabulary, isolated word, speaker dependent or multi-speaker speech recognition due to its relative simplicity and good recognition performance in these situations. Moreover, DTW approach is limited to word template (Ursin, 2002).

Although DTW is computationally much simpler than HMM, its requirement is still quite high. HMM can capture the statistical characteristics of word and sub-word units among different speakers even in large vocabulary and thus, it is better than DTW in speaker independent large vocabulary speech recognition (Wong, 1998).

3.5.2 The Hidden Markov Models (HMMs)

In most current speech recognition systems, the acoustic modeling

components of the recognizer are almost exclusively based on Hidden Markov Model (HMMs). The ability to statistically model the variability in speech has been the main reason for the success that HMMs have enjoyed over years (Ming, 2007).

HMMs provide an elegant statistical framework for modeling speech patterns using a Markov process that can be represented as a state machine. The temporal evolution of speech is modeled by an underlying Markov process. The probability distribution associated with each state in an HMM models the variability which occurs in speech across speakers or even different speech contexts (Ming, 2007).

Basically, HMM consists of a number of states; each state having its own mapping of the feature space into an observation probability space. The observation probability for a feature vector therefore depends on the state.

3.5.3. Artificial Neural Networks (ANN)

Artificial neural networks (ANN) are a powerful and flexible architecture for solving classification problems. ANNs have also been applied to speech recognition owing to several advantages they offer over the typical HMM systems. ANNs can learn very complex non-linear decision surfaces effectively and in a discriminative fashion. Besides, they are easy to implement(Podder,1997).

The discriminating power of neural nets is their main advantage over HMMs. For speech applications, however, neural nets have so far been unable to match the performance of HMM based systems because of the HMM's

superior modeling of temporal structure whereas the ANNs are typically formulated as classifier of static data. Moreover, the computation needed for the estimation process of ANNs is significantly more expensive than HMMs(Podder,1997).

3.6 Problems with speech recognition

Although Substantial progresses have been made in the area of automatic speech recognition, many problems related to speech recognition still exist. A National Science Foundation survey in the United States has identified key research challenges related to human language technology (Cole et. al, 1995). The following are some of the challenges identified with regards to automatic speech recognition.

3.6.1 Robustness

The current most advanced speech recognition systems can achieve very high accuracy rates if trained for a particular speaker, in a particular language and speaking style, in a particular environment and limited to a specific task. It remains a serious challenge however to design a recognition system capable of understanding virtually anyone's speech, in any language, on any topic, in any style in all possible environments. Thus, a speech recognition system trained in a laboratory with clean speech may degrade significantly in the real world due to noise in the input speech signal.

Robustness in speech recognition refers to the need to maintain good recognition in spite of variability in speech signal.

3.6.2 Variability in Speech signal

Automatic speech recognition is often viewed as a mapping from the speech signal to a sequence of discrete entities such as phonemes, words, and sentences. A major obstacle in obtaining high-accuracy recognition is the large variability in the speech signal characteristic. As described in the work of Makhoul and Schwartz,(1995), there are three components of speech variability. These are, linguistic variability, speaker variability, and channel variability.

Linguistic variability: includes the effects of phonetics, phonology, syntax, semantics, and discourse on the speech signal. When a word with different meaning has the same phonetic realizations, it becomes impossible for a recognition system to find the correct sequence of words.

Speaker variability: refers to the fact that every individual speaker is different and will have a unique speech pattern. This is known as inter-speaker variability. Speakers can differ in a number of different ways. Individual's physical constitutions such as age, health, size, lung capacity. The size and shape of the vocal tract determines the characteristics of his/her speech. Gender also plays an important role as the pitch of a female voice is in general significantly higher than male speakers.

Dialect speakers use different phonemes to pronounce a word than non-dialect speakers or use different allophones of certain phonemes. Further, social environment, education and personal history also affect the manner in which individuals speak.

Intra-speaker variability is also described by the fact that even the same person can not be able to exactly reproduce an utterance. This can be caused by a number of reasons such as fatigue and difference in emotional state (irritated, frustrated). When speaking in a noisy environment a person's voice also tends to differ. This is called the Lombard effect. In general, in such circumstances vowels will grow longer and much of the voice's energy will shift to higher frequencies.

Channel variability: includes the effects of background noise and the transmission channel such as microphone and telephone effect on recorded or transmitted sound.

Furui (1995) classified the main causes of speech variation based on whether they originate in the speaking and recording environment, the speakers themselves, or the input equipment, indicated in Table 3.1. Additive noises can be classified as stationary or non stationary, with most typical non stationary noise being other voices. In addition noise can be classified according to whether they are correlated or uncorrelated to speech.

Environment	**Speaker**	**Input Equipment**
Speaker correlated noise- reverberation, reflection	***Attribute of speakers***-dialects, gender, age	***Microphone(transmitter)-*** Distance to microphone
Uncorrelated noise- additive noise both stationary and non-stationary	***Manner of speaking-*** breath, lip noise, stress, rate, level, pitch, cooperativeness	***Filter Transmission system-*** distortion ,noise ,echo Recording equipment

Table 3.1: Main causes of speech variation (adopted from Furui, 1995)

3.7 Review of Related Works

There are various attempts in the area of Speech recognition for Ethiopian as well as non Ethiopian languages. Reviewing researches conducted in these area helps to find out approaches and techniques that are ultimately helpful to conduct this research. Accordingly, the efforts that have been made so far are reviewed to provide some insights for this research.

Solomon (2001) has attempted the possibility of building speaker dependent and speaker independent HMM based isolated Consonant-Vowel (CV) syllable recognition for Amharic. He argues that syllable has a close connection to articulation and hence, it has a potential for a relatively compact representation of conventional speech. He has also used a bi-gram language models trained using Hidden Markov Toolkit (HTK) and observed high perplexities due to inflectional and derivational morphological feature of Amharic. The designed speech recognizer prototype achieved 90% word accuracy, which is considered to be promising.

As conclusion, he proposed the CV-syllables to be the best candidates for the basic recognition units for Amharic. However, the problem of germination of consonants was not incorporated in the pronunciation dictionary.

Solomon's work was extended by Kinfe (2002) who used the HTK Toolkit to build HMM word recognizers at three different sub-word levels: phoneme, tied-state triphone, and CV-syllable. He collected a 170 word vocabulary from 20 speakers. He considered a subset of the Amharic syllables, concentrating on the combination of 20 phonemes with the seven vowels, or

in total 140 CV-units. His training and testing sets both considered 50 discrete words. Having this corpus, he attempted to construct phoneme based, triphone based and CV syllable based speech recognizers with both speaker dependent and speaker independent models. Contrary to Solomon's prediction, the performance of the CV syllable recognition was very poor.

Kinfe presented an isolated word recognition accuracy of 83.1% for phoneme based and 78.0% for triphone based speaker dependent models while the speaker independent models gave 75.5% accuracy for phone-based and 77.9% for triphone based isolated word models. Hence, he recommended tied state triphone as a good sub-word unit for Amharic recognizers.

Zegaye,(2003) further investigated the possibility of developing large vocabulary, continuous speech recognition and speaker independent Amharic Speech recognizers. Zegaye tried to build phone based and triphone based recognizers for Amharic using Hidden Markov Toolkit (HTK). For the purpose of training and testing the recognizer, he used a speech data previously recorded by Solomon (Solomon, 2001). Accordingly, he indicated that a triphone based recognizer with 76.2% word accuracy achieved a better performance.

Ashenafi(2009) studied the possibility of developing an automatic speech recognizer for Afaan Oromo. He built an isolated word speaker independent speech recognizer using a Hidden Markov Model (HMM) and an open source recognition tool called sphinx.

Ashenafi (2009) attempted to construct a speech corpus of 1000 datasets containing 50 selected Afan Oromo words and uttered by 20 different people. The experimentation involved context independent word based models and content dependent phoneme based models. To evaluate the performance of the recognizer models, the recognition accuracy, the transcription speech, and types of errors occurred in decoding the test sets were measured and compared.

Accordingly, his work has presented 82.83% and 81.081% word level accuracy for context dependent phoneme based models and context independent word based models, respectively.

Hafte (2009) tried to build a large vocabulary continuous speech recognition system for Tigrigna using statistical approach. To build the speech recognizer, a speech corpus composed of 250 utterances for the training and 50 sentences for testing the models were constructed.

Hafte constructed a bi-gram language model to support the acoustic model. The study also used phonemes and triphones as units of recognitions. The recognition models were tested and their respective performances were evaluated. In the end, the evaluation of speaker independent models achieved a 60.20% word level correctness, 58.97% word accuracy, and 20.06% level of correctness. The results of the experiments also indicated that phonemes and tied state triphones came up with better performance. Hence, both units can be best candidates for basic recognition units for Tigrigna.

CHAPTER FOUR

SPEECH RECOGNITION MODEL

Processes from the real world usually produce outputs that can be observed, and these outputs are characterized as signals. Signal-like characters from an alphabet and quantized vectors from a codebook can be discrete. Alternatively, signals such as speech wave, temperature values and music can be continuous. Signal can also be classified as stationary or non-stationary. It can be pure, noisy, or corrupted by transmission of distortions and reverberation (Rabiner, 1989).

In order to represent these real world signals into major parts based on their characteristic features signal models are used. The signal models can be categorized into two parts, deterministic models and statistical models. Deterministic models deal with some known specific properties of a given signal where as statistical models focus on only the statistical features of the signal (Rabiner,1989).

One of the most flexible and successful statistical model to speech recognition so far is the Hidden Markov Models (HMMs). In this section basic concepts of HMMs are presented, the algorithms for training and using them are described, and problems associated with HMMs are also reviewed. Finally, the Sphinx system is briefly explained.

4.1. The Hidden Markov Models (HMM)

Hidden Markov Models (HMMs) are the most predominant and widely used statistical models in characterizing speech. HMMs were first described in a classic paper by Baum (Baum, 1972). Later on it was extended to ASR independently at CMU and IBM (Ming, 2007).

The major reason most modern recognition procedures emphasize on HMM based recognition is that, it has a rich mathematical structure to provide better results for typical large vocabulary continuous recognition systems.

According to Rabiner(Rabiner,1989) this approach is popular because of the precise mathematical framework, the ease and availability of training algorithms for estimating parameters of the models from finite training sets of speech data.

A Hidden Markov Models (HMM) has two components, an underlying Markov chain having a finite number of states and a finite set of output probability distributions, one of which is associated with each state. Hence, it is extremely useful for modeling sequentially changing behavior as in speech recognition application (Levinson et al.,1988).

When an HMM is applied to speech recognition, the states are interpreted as acoustic models, indicating what sounds are likely to be heard during their corresponding segments of speech; while the transitions provide temporal constraints, indicating how the states may follow each other in sequence. Because speech always goes forward in time, transitions in a speech application always go forward or make a self-loop, allowing a state to have

arbitrary duration. Hence states and transitions in an HMM can be structured hierarchically, in order to represent phonemes, words, and sentences (Tebelskis, 1995).

HMM is also described as a finite-state machine that changes state once every discrete time. At discrete instants of time, the process is assumed to be in some state and an acoustic speech vector or observation is generated by the random function corresponding to current state. The underlying Markov chain then changes states according to its transition probability matrix. Only the observation of the acoustic speech vector is known, but the underlying Markov chain is hidden, hence, it is called hidden Markov model(Magdi & Gader, 2000).

HMMs represent speech as a sequence of observation vectors derived from a probabilistic function of a first-order Markov chain Model. States are identified with an output probability distribution that describes pronunciation variations, and states are connected by probabilistic transitions that capture durational structure. HMM can thus be used as a maximum likelihood classifier to compute the probability of a sequence of words given a sequence of acoustic observations (Hansen,2003).

HMM generates a discrete time random process consisting of two sequences of random variables which are the unseen (hidden) states and the known observations. The underlying structure of an HMM is the set of states and associated probabilities of transitions between states known as a Markov

chain. Figure 4.1 shows a 3-state Markov chain with transition probabilities (Hansen, 2003).

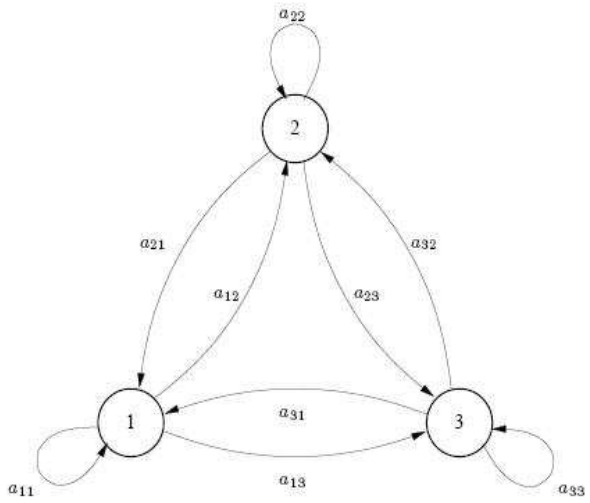

Figure 4. 1: A 3-State Markov Chain with Transition Probabilities (Hansen, 2003)

In Figure 4.1 the Hidden Markov Model is described as a collection of states connected by transitions. It begins in a designated initial state. In each discrete time step, a transition is taken into a new state, and then one output is generated in that state. The choice of transition and output symbol are both random, governed by probability distributions. Hence, the HMM can be thought of as a black box, where the sequence of output symbols generated over time is observable, but the sequence of states visited over time is hidden from view(Tebelskis,1995).

The main elements of HMM include the number of states in the model, the number of distinct observation symbols per state, the state transition probability distribution, the observation symbol probability distribution and the initial state distribution (Rabiner, 1989;Rabiner and Juang, 1993).

Rabiner (1989) stated two strong reasons for the importance of HMMs. Firstly, HMM models are very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of applications. Secondly, HMM models work very well in practice for several important applications provided that they are applied properly.

4.2. Mathematical Representation of HMMs

Currently, the most successful method of speech recognition is based on statistical pattern recognition. The objective of any statistical based speech recognizer is to find the most likely word sequence from list of words that could possibly be generated given the acoustic data (Helin, 1999).

In the context of HMMs, the speech recognition problem is decomposed as follows. Speech is broken into a sequence of acoustic observations or frames, each accounting for around 25 milliseconds of speech; taken together, these frames comprise the acoustics associated with an utterance. In this section the mathematical representation of HMM based speech recognizer is explained based on the works of Jelinek (Jelinek, 1997) as follows:

If we consider a simple probabilistic model for speech production with W= w1, w2,..., wn as the sequence of specified words that are modeled and A as the sequence of acoustic evidence or observation that is provided to the system, then the recognition system which is based on the stochastic models must select a word string W* that maximize the probability that the word string W was spoken given that the acoustic observation A was observed is shown as:

$$W^* = \underset{w}{\operatorname{argma}} P(W|A) \tag{4.1}$$

Where, P(W|A) is termed as the a posteriori probability since it represents the probability of occurrence of a sequence of words after observing the acoustic observation A. This is known as the Maximum A Posteriori or MAP criterion. The above approach to speech recognition, where the word hypothesis is chosen within the probabilistic framework, is what makes most present speech recognizers statistical pattern recognition systems (Helin, 1999).

However, the above equation is not directly computable since the number of possible observation sequences are infinite for a given language from which most likely word sequence needs to be chosen. This problem can be significantly simplified by applying the Bayesian approach to find W*.

Hence, using Bayes' formula we can rewrite the conditional probability to be maximized as:

$$\mathbf{W^*} = \underset{w}{\operatorname{argmas}}_W \quad \frac{P(A|W)P(W)}{P(A)} \tag{4.2}$$

Where P(W) is the probability that the word string W will be uttered and P(A|W) is the probability that when the word string W is uttered the acoustic signal A will be produced . Finally P(A) is the probability that A will be observed. Since A is fixed- the acoustic evidence is already given and does not change through the recognition process- the recognition problem can be decomposed into two parts as the language modeling problem and the acoustic modeling problem as described below. Hence, we can discard P(A)

and rewrite equation (4.2)

as:

$$W^* = \underset{W}{\operatorname{argmax}}\ P(W)\,P(A|W) \tag{4.3}$$

In practice, only the acoustic vectors are known and the underlying word sequence is hidden. The probability P(A|W) is evaluated by the acoustic model which estimates the probability of the acoustic observations. The probability of the word sequence P(W) is evaluated using the Language model(Kim,2004). So the word string W* that maximizes the product of P(W) and P(A|W) is recognized by the speech recognizer system.

The a priori probability is typically given by a language model in speech recognition system. P(A) is the probability of the acoustic data, which is independent of all word sequence hypotheses, so that it can be ignored. Probabilities for word sequences are generated as a product of the acoustic and language model probabilities. The process of combining these two probability scores and sorting through all plausible hypotheses to select the one with the maximum probability is called decoding or search (Ming, 2007). Figure 4.2 depicts the above framework under which most ASR systems operate.

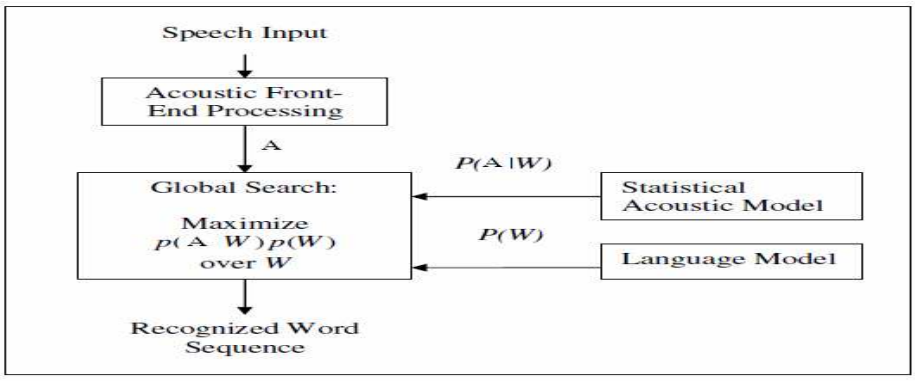

Figure4.2. Schematic representation of Statistical Speech recognition (adopted from: Ming, 2007)

Consequently, a speech recognizer consists of three components: acoustic Front-End preprocessing part that translates the speech signal into a sequence of observation symbols, a language model that tells us how likely a certain word string is to occur, and a statistical acoustic model that tells us how a word string is likely to be pronounced. (Ming, 2007).

4.2.1. Probability of a Sentence

P(W) defined above can be viewed as "the probability of a sentence". Actually we need this notion in order to select the string of words which maximizes $P(W|A)$. Hence, Probability of a word string can be formulated as follows:

$$P(W)P(w1,w2,w3,...wn) \tag{4.4}$$

Using the chain rule:

$$P(W)=P(w1)P(w2/w1)p(w3/w_1^{2}),...,P(w_n/w_1^{n-1}) \tag{4.5}$$

$$P(W) = \prod_{k=1}^{n} P(w_k / w_1^{k-1}) \tag{4.6}$$

Where w_1^{k-1} is the sequence of words preceding w_k. It is the history of w_k. If there were no dependencies between the words; i.e. occurrence of a word did not depend on the previous word sequence w_1^{k-1} then we could compute the probability of W as:

$$P(W)=P(w_1),P(w_2),P(w_3),... ,P(w_n) \tag{4.7}$$

Where $P(w_k)$ is the probability of the occurrence of the word w_k. The assumption that each word is uttered independently employs a "unigram model". The formulation of the assumption is as follows:

$$P(W_k/ w_1^{k-1}) = P(w_k) \quad (4.8)$$

The occurrence of a word depends on its history by means of syntax and semantic context. The range of this dependency can go beyond to the beginning of a sentence or utterance, even to the beginning of a speech or conversation. Unfortunately there is no known way to formulate this dependency for every string of words in both linguistic and probabilistic terms. Hence, this assumption is not valid for continuous speech. The occurrence of a word depends on its history by means of syntax and semantic context.

The solution is to limit the range of dependency to N words and define the history of a word as the previous N-1 words. In other words the following assumption is made:

$$P(W_k/ W_1^{k-1}) \approx P(W_k| W_{k-1}, W_{k-2}, \dots, W_{k-N+1}) \quad (4.9)$$

The language models based on this assumption are called "N-gram models", and the word sequences W_{k-N+1}, W_{k-N+2},...,W_K are called N-grams. So, a bi-gram model assumes that:

$$P(W_k/W_1^{k-1}), P(W_k|W_{k-1}) \quad (4.10)$$

and a trigram model assumes:

$$\mathbf{P(}W_k/W_1^{k-1}\mathbf{)}, P(W_k|W_{k-1},W_{k-2}) \quad (4.11)$$

Where, W_{k-1}, and W_{k-2} are the two words just before the word in question.

4.2.2. Elements of Hidden Markov Models (HMM)

The basic elements of a hidden Markov Model are discussed below.

- The number of states N. Although the states are Hidden, for many practical applications there is often some physical significance attached to the states or to sets of states of the model (Rabiner and Juang,1993). The individual states are represented by:

  N = {1, 2, 3, . . . , N}, and a state at a time t is denoted by qt

- Model parameter M.
- Initial state distribution $\pi = \{\pi_i\}^N{}_{i=1}$ in which π_i defined as:

$$\pi_i = P(q_1=1) \tag{4.12}$$

- state transition probability distribution A= (aij)= where:

$$aij= P(q_t+1= j|q_t = i), \quad 1<=i,j<=N \tag{4.13}$$

- Observation symbol probability distributions, $B = \{b_j(O_t)\ N_j=1$, in which the probabilistic function for each state, j ,is:

$$b_j(O_t) = P(O_t\ q_t = j) \tag{4.14}$$

The calculation of bj (Ot) can be found with continuous or Discrete observation densities used. A complete specification of an HMM requires two model parameters M and N as well as the specification of three set of probability measures, π, A, B. where:

$$\lambda = (\pi, A, B) \tag{4.15}$$

4.2.3 Assumptions in the theory of HMM

For the sake of mathematical and computational tractability, the following assumptions are made in the theory of HMMs.

4.2.3.1. The Markov assumption

As given in the definition of HMMs, transition probabilities are defined as,

$$\{aij = p\{q_{t+1} = j | q_t = i\} \tag{4.16}$$

In other words it is assumed that the next state is dependent only upon the current state. This is called the Markov assumption and the resulting model becomes actually a first order HMM. However generally the next state may depend on past k states and it is possible to obtain such a model, called k^{th} order HMM by defining the transition probabilities.

But it is seen that a higher order HMM will have a higher complexity. Even though the first order HMMs is the most common, some attempts have been made to use the higher order HMMs as well.

4.2.3.2 The Stationary Assumption

Here it is assumed that state transition probabilities are independent of the actual time at which the transitions take place. Mathematically,

$$P\{ q_{t1+1} = j \mid q_{t1} = i\} = P\{ q_{t2+1} = j \mid q_{t2} = i\} \tag{4.17}$$

4.2.3.3. The output independence assumption

The out put independence assumption implies that current observation is statistically independent of the previous observations. We can formulate this assumption mathematically, by considering a sequence of observations as:

$$O = o_1, o_2, \ldots o_T$$

Then according to the assumption for an HMM λ,

$$p\{ O \mid q_1, q_2, \ldots, q_T, \lambda \} \prod_{t=1}^{T} p(o_t \mid q_t, \lambda). \quad (14.18)$$

However unlike the other two, this assumption has a very limited validity. In some cases this assumption may not be fair enough and therefore becomes a severe weakness of the HMMs.

4.3. The three Problems of HMM

For Hidden Markov Models to be useful in building speech recognizers there are three key problems of interest that must be addressed in order to apply HMM into the real applications. These problems are described based on the works (Juang & Rabiner, 1991; Rabiner, 1989).

4.3.1. The First Problem: Evaluation Problem

The first problem can be described as follows:

Given the observation sequence $O = (O_1, O_2, \ldots, O_T)$ and the model $\lambda = (\pi, A, B)$, how do we efficiently compute $P(O \mid \lambda)$, the probability of the observation sequence, being produced by the model?

This problem is viewed as getting a score on how well a given model or a sequences of models matches a given observation sequence corresponding the speech to be recognized. This is useful when we need to choose among several competing models, the solution of the problem allows us to choose the model which best match the observations for the purpose of classification and recognition (Rabiner, 1989).

4.3.2. The Second Problem: Decoding Problem

The second problem can be formulated as follows:

Given the observation sequence $O = (O_1, O_2, ..., O_T)$ and the model λ, how do we choose the most likely state sequence $q = (q_1, q_2, ..., q_T)$ according to some optimality criterion ,for example, it is most suitable to explain the observations?

The second problem is decoding problem, in which we attempt to uncover the most likely state sequences that lead to the observation sequences. There are several optimality criterions that can be imposed. Hence the choice of criterion is a strong function of the intended use for the uncovered state sequences.

Typical use is to find the optimal state sequences for continuous speech recognition. Another use is if we use the states of a word model to represent the distinct sounds in the word, it may be desirable to know the correspondence between the speech segments and the sounds of the word, because the duration of the individual speech segments provides useful in formation for speech recognition (Rabiner, 1989).

4.3.3. The Third Problem: Estimation Problem

The estimation problem can also be described as follows:

Given the observation sequence $O = (O_1, O_2, ..., O_T)$, how to adjust the model parameters $\lambda = (\pi, A, B)$ such that $P(O \mid \lambda)$ is maximized?

Given an observation sequence or a set of sequences, O, the estimation problem involves finding the best model parameter values that specify a model most likely to produce the given sequence. In speech recognition, this is often called training, and the given sequence, on the basis of which we obtain the model parameters, is called the training sequence. The training problem allows us to optimally adapt model parameters to the training data to create the best model for the phenomena (Rabiner, 1989).

4.4. Hidden Markov Model Algorithms

In order to deal with previously stated problems HMM possesses three elegant and efficient algorithms. These algorithms provide solutions to the previously discussed three HMM problems that should be addressed in order to apply HMM models to real world applications. Based on the works of (Podder,1997), Rabiner and Juang (1993) and Rabiner,1989), these algorithms are discussed below.

4.4.1. Forward-backward Algorithm

The forward-backward algorithm provides solution to the first problem described above. It is an algorithm for computing the probability of a particular observation sequence in the context of hidden Markov models. The algorithm first computes a set of forward probabilities which provide the probability of observing the first k observations in the sequence and ending in each of the possible Markov model states.

The algorithm also computes a set of backward probabilities which provide the probability of observing the remaining observations given an initial state.

These two sets of probabilities can then be combined to provide the probability of being in each state at a specific time during the observation sequence. The forward-backward algorithm can thus be used to find the most likely state for a hidden Markov model at any time.

4.4.2. Viterbi Algorithm

Viterbi algorithm provides a solution to the second Problem described above. It is an inductive procedure in which at each instant of time it keeps the best (i.e. the one giving maximum probability) possible states sequence for each of the N states as the intermediate state for the desired features vector sequence Z. It finally yields the best path for each of the N states as the last state for the desired observation sequence. Out of these, the one which has the highest probability is selected.

So, the gist of the Viterbi algorithm is that it estimates the optimum states sequence, and it does so by reformulating the problem accurately. It is a maximization of the Forward-Backward algorithm.

4.4.3 Baum-Welch algorithm

Baum-Welch algorithm provides a solution to the third Problem, which is a maximization of the solution to the first Problem through the adjustment of the parameters of the model. This optimization criterion is called the maximum likelihood criterion.

To determine the parameters of a HMM it is first necessary to make a rough guess at what they might be. Once this is done, more accurate (in the

maximum likelihood sense) parameters can be found by applying the so called Baum-Welch re-estimation formula. Since the full likelihood of each observation sequence is based on the summation of all possible states sequences, each observation vector (features vector) contributes to the computation of the maximum likelihood parameter values for each state.

In other words, each observation is assigned to every state in proportion to the probability of the model being in that state when the vector is observed. Such probability is efficiently calculated using the Forward-Backward algorithm.

Since the full likelihood of each observation sequence is based on the summation of all possible states sequences, each observation vector (features vector) contributes to the computation of the maximum likelihood parameter values for each state. In other words, each observation is assigned to every state in proportion to the probability of the model being in that state when the vector is observed. Such probability is efficiently calculated using the Forward-Backward algorithm.

4.5 Types of Hidden Marcov Models

The most commonly used types of HMMs are described as follows.

4.5.1 Fully Connected or Ergodic HMM

The Ergodic HMM is the most general structure used for modeling HMM. In this model any state can be reached from every other state of the model. In order to fulfil the ergodic model, the property, $0 < aij < 1$ has to be maintained. The state transition matrix, A, for an ergodic model can be

described by:

$$A = \begin{bmatrix} a11 & a12 & a13 & a14 \\ a21 & a22 & a23 & a24 \\ a31 & a32 & a33 & a34 \\ a41 & a42 & a43 & a44 \end{bmatrix}$$

$4x4$

4.5.2. Left-Right HMM Model

In speech recognition, is desirable to use a model which models the observations in a successive manner- since this is a property of speech. The model that satisfies this modeling technique is the Left-Right model or Bakis Model(Markowitz,1996). The fundamental property of all Left-Right HMM is that the state transition coefficient should have the property:

$$aij = 0, \text{ where } j < i$$

That is, no jumps can be made to the previous states. The length of transitions is usually restricted to some maximum length, typically two or three:

$$aij = 0, \text{ where } j > i + \Delta$$

The state transition matrix, A, for the left-right model when $\Delta = 2$, is presented by:

$$A = \begin{bmatrix} a11 & a12 & 0 & 0 \\ 0 & a22 & a23 & 0 \\ 0 & 0 & a33 & a34 \\ 0 & 0 & 0 & a44 \end{bmatrix}$$

4x4

The choice of constrained model structure like a Left-Right model requires no modification in training algorithms (Rabiner,1989). This is because any state probability set to zero remains zero in the training algorithms.

4.6. Speech Recognition Tools

Automatic Speech Recognition is a technology that enables a computer to identify the words that a person speaks in microphone or telephone (Satori et. al, 2004). Speech recognition systems applications have been used in areas such as command recognition, dictation, Interactive Voice Response. It can also be used to learn new languages and for disabled people to interact with societies through the help of computers.

4.7 The Sphinx Systems

To highlight the importance of speech recognition systems in a society a number of recognition systems have been developed. The most commonly used are Dragon naturally speaking, IBM via voice and Microsoft SAPI. The most popular Open source speech recognition tools are HTK, ISIP, AVCSR, and CMU Siphinx-4(Satori et. al, 2004).

Sphinx-4 is an open source HMM-based speech recognition system written in the Java programming language (Lamere et.al, 2003). The Sphinx-4 decoder is designed jointly by researchers from CMU, Sun Microsystems Laboratories, and Mitsubishi Electric Research Laboratories.

According to Lamere et.al.,(2003), over the past few years, the demands placed on conventional recognition systems have increased significantly. However, several functionalities are now additionally desired of a system

such as ability to perform multistream decoding with much user control and some degree of easy control over system's performance in the presence of unexpected noise, portability across growing computational platforms, conformance to widely varying resource requirements, easy restructuring of the architecture for distributed processing and good and flexible user interface.

The design of Sphinx-4 is also driven by almost all of these considerations, resulting in a system that is highly modular, portable and easily extensible. Many new features have been incorporated by extending conventional design strategies or implementing new ones. The design is more utilitarian and futuristic than most existing HMM-based systems.

4.7.1. General Architecture of Sphinx-4 Speech Recognizer

The Sphinx 4 has been designed with a high degree of modularity. Figure 4.3 shows the overall architecture of the system. The main blocks are Front end, Decoder and Knowledge base. Except for the blocks within the KB, all other blocks are independently replaceable software modules written in the Java™ programming language. Stacked blocks indicate multiple types which can be used simultaneously (Lamere et. al, 2003).

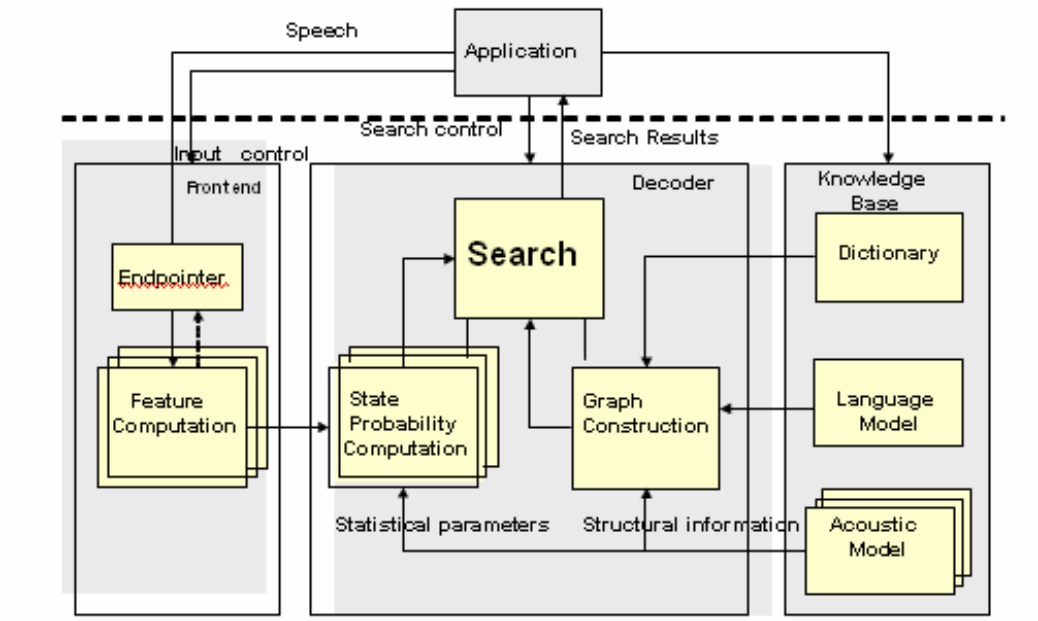

Figure 4.3: General architecture of Sphinx system (Source: Lamere et al. ,2003)

The main blocks in Siphnix-4 architecture are front-end, decoder and knowledge base (Walker et al., 2004) and the function of the components are discussed as follows.

4.7.1.1. The Front-end component

The purpose of the Font-end is to parameterize an input signal(i.e speech) in a sequence of output features. As illustrated in figure 4.2,the Font-end comprises one or more sequential chains of replaceable communicating signal processing modules called Data processors. Supporting multiple chains permits simultaneous computation of different parts of parameters from the same input signal. This enables the creation of systems that can simultaneously decode using different parameter types.

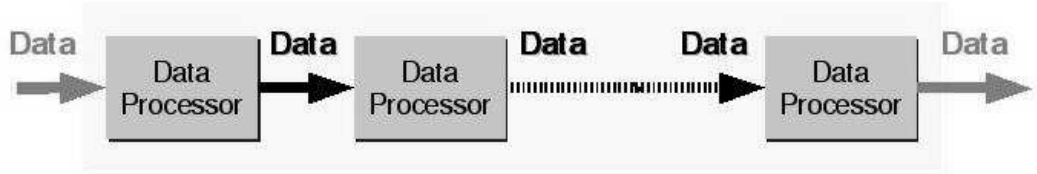

Figure 4.4: The Sphinx 4 Font-end (Source: Lamere et al.,2003)

The Front-end module is made up of several communicating blocks. Each communicating block consists of an input and an output. Each input is connected to the output of its predecessor and interprets it to find out if the incoming data is a speech data or a control signal, which can show the beginning or end of speech.

This design allows the system to be used in live mode, and allows each output to be read. Actual input to the system does not have to be at the first block, but can be at any of the blocks. Sphinx-4 allows users to run the system using speech signals, different bispectra, spectra, cepstra and others. Thus, other kinds of features such as auditory representations can be plugged in. Additional blocks can also be put between any blocks (Lamere et. al, 2004).

Based on the works of Lamere et. al. (2004) the Sphnix-4 systems have four modes of operation and are discussed as follows. First, when a system is running continuously from a stream of input speech. Second, when a system performs end pointing (determining both beginning and ending endpoints of a speech segment automatically). Third, the user gives the beginning of a speech segment but the system determines when the speech ends automatically. Finally, the fourth mode is when the user gives both the

beginning and the end of a speech segment.

End point detection is performed by an algorithm that compares the three energy threshold levels. Two of these are used to describe when speech is starting and one for determining end of speech. The end pointer detects the start and end of speech, non-speech segments, which are not sent to the decoder.

4.7.1.2. The Sphinx Decoder

Decoder is a Software program that takes the sounds spoken by a user and searches the Acoustic Model for the equivalent sounds. When a match is made, the Decoder determines the phoneme corresponding to the sound. It keeps track of the matching phonemes until it reaches a pause in the users' speech. It then searches the Language Model or Grammar file for the equivalent series.

4.7.1.3. Sphinx Knowledge Base

The Knowledge base also called the Linguistic provides the information the decoder needs to do its job. According to Satori et al.(2003), the knowledge base of a Sphinx system is made up of three components. These are the Acoustic model, Dictionary and the language model.

The acoustic model contains statistical representation of the sounds created by training using enough acoustic data. The dictionary contains list of words with their respective pronunciation information. Where as, the language

model contains statistical representation of the probability of occurrence of words in the system.

4.7.2 The SphinxTrain

SphinxTrain is the acoustic training environment for all versions of CMU Sphinx. It contains suitable programs, documentations and scripts for building acoustic models for any language as long as there is enough acoustic data. The importance of SphinxTrain can be explained based on the fact that it is impossible to have a functional recognizer without building the acoustic model which is necessary for comparing the data coming from Front End Module (Satori et al., 2003).

CHAPTER FIVE

EXPERIMENTATION AND PERFORMANCE ANALYSIS

This chapter describes the design, implementation and performance analysis of the prototype Sidaama speech recognition system. The main goal is to test the suitability of a new language model for a speech recognizer of Sidaama. In this research two recognizer models were constructed. The first context dependent triphone based isolated word recognizer. The second is context independent monophone based isolated word recognizer. Finally, based on the test speech datasets used their respective performances are evaluated and presented.

The main objective of the recognizers' performance evaluation undertaken in this research is to investigate the extent to which the recognizers work, based on which to report findings and forward recommendations for further study in the area.

Basically, the two built recognizers utilized a unigram language model in order to test and compare the performances of both CI monophone based models and CD triphone based models and hence forward recommendations as per their outcomes.

The development process is performed on a Microsoft windows operating system platform using more flexible and state of the art speech recognizer with appealing features called sphinx-4. The discussion of the experiment is presented in accordance to the steps followed in building the prototype Sidaama speech recognizers.

5.1 The Speech Recognizer Design

Clearly putting a system design that shows the whole processes of the system and the interaction among components provides benefits for the system users as well the system developers. Moreover, systems have a number of components that interact with each other in logical fashion. The processing logic of a component may also generate an output that may be the input for another component in the system. Similarly, it may also trigger another component to activate itself or it may affect the process logic of a particular component. Hence, taking these design issues into account the Sidaama speech recognizer prototype model is developed. The prototype design shows the various system components, interactions and the training and recognition phases of the recognition process using HMM.

As shown in the figure 5.1, the recognizer design has training, decoding and testing phases. Sidaama speech corpus along with the corresponding text corpus is used as the main input for the speech recognizer.

During the training phase, activities such as manual transcription of the text corpus as well as feature extraction from the speech corpus were performed to create the recognition grammar and the HMMs models outputs. In turn, the recognition grammar and the HMMs are used as an input for the next phase of the speech recognition process, the decoding. In this phase, the recognizer lexicon, the acoustic model and the language model are integrated together through the help of the recognizer to decode an input speech utterance into a text word.

Finally, to test the performance of the prototype speech recognizer, accuracy tracker component is integrated in to the recognizer design. This accuracy tracker accepts input speech utterances together with their manual transcription to test how correctly the recognizer identifies the words.

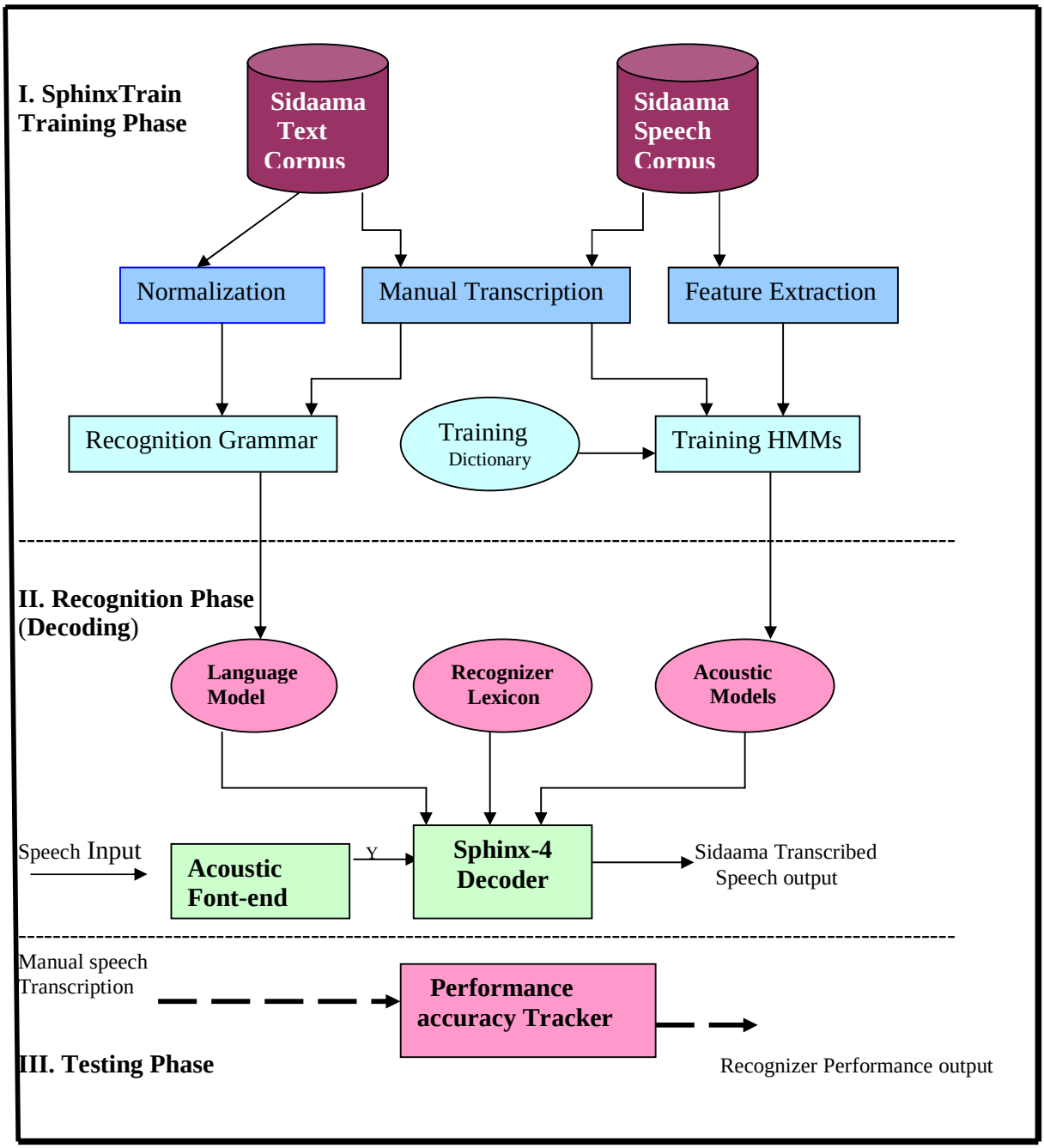

Figure 5.1: Overview of HMM based ASR design for Sidaama

5.2. Experimentation of the Prototype Recognizer Models

The experimentation task involved data preparation which is recording and preprocessing activities, building the language model for speech recognizers model, training the HMM or creation of the acoustic models and finally building a decoder that coordinates the language model, the pronunciation dictionary and the acoustic model in order to evaluate and present the performance of the recognizer models.

5.2.1 Data Recording and Preprocessing

The first task in the experimentation of this research is preparing the data necessary to construct the acoustic models. Right before going on to the task of recoding and making the utterances available for experimentation Meta data was prepared. Schiel and Draxler(2004) in their work described three types of header information that most speech corpora contain as a meta data. These are speaker profiles, protocols used during recording, and diverse comments. Accordingly, a Meta data for producing the speech corpus consists the session ID, the record date, and the environmental and technical recording conditions.

The session ID is used to identify a single particular speech data from a given speech corpus. It consists of elements that give broad categorical information about the recording such as the language, sex of the speaker, the type of the recording setup, and the identification number. Hence, the session ID is often used in speech corpus as a file name to identify speech data belonging to the same category. For instance, one particular session ID of a recording is S1001, indicating 'S' Sidaama, and '1001' is a unique ID.

Regarding the environmental condition as it has been described in (section 1.5) a quite office room is selected in order to maintain the quality of utterance being recorded.

Each of the 450 words is recorded during the utterance of a single speaker. These words were selected upon consultation of a domain expert and commented to be phonetically balanced to represent the sounds in the language. Samples of selected distinct words for the experiment are attached in Appendix III.

Having the recorded speech utterances, 1/3 by 2/3 dataset division is made to form the test set and training set respectively. Table 5.1 shows the amount of speech corpus classified as training set and test set.

Data sets	**Words used for the research**
Training	300
Test words sleeted from trained corpus	100
Test words not used for training	150

Table 5.1: Proportion of speech corpus used for the study

5.2.2 The Sidaama Phoneme Sets

Once sorted list of words that are included in the dictionary are identified, there is a need to find phonemes of each word. The phoneme sets used in this research are generated using Sphinx knowledge base tool[2]. However, the

[2] sphinx Knowledge base tool is available online "http://www.speech.cs.cmu.edu/tools/lmtools- adv.html

phoneme sets generated from this tool are not directly applicable for modeling speech recognition task for the language under study. This is because the script used is not specifically designed for Sidaama language. Hence, for phonemes that are incorrectly mapped a manual modification is performed. For instance, the tool generates the phoneme set {K AA K OW} for the word CAQQO which should be actually transcribed as {C AA QQ OW} since C has the sound /tʃ'/ not /k/.

Some of the words that are wrongly generated using the online Sphinx Knowledge base tool and needed manual transcription are presented in table 5.2.

Sidaama Words	**wrongly extracted phonemes using the script**	**Manually modified words**
CAAQQO	K AA K OW	C AA QQ OW
DHAANSA	D HH AA N S AH	DH AA N S AH
AMANYOOTE	AH M AE N AY UW T	A M AA NY OO T EH
XUMA	Z UW M AH	X UW M AA
MAXAAFA	M AE K S AH F AH	M A X AA F A

Table 5.2: Sidaama special phonemes and their transcription

Special phonemes that appear in Sidaama language which have been wrongly extracted from the text corpus when the tool is used are described in the table 5.3 below.

C c	DH dh	NY ny	X x
[ʧ']	[ð]	[ɲ]	[tˀ]

Table 5.3: Special phonemes in Sidaama language

Once the words in the text corpus are transcribed to the corresponding phonemes that describe the language, each unique phoneme was collected to form the total phoneme sets for Sidaama language. In addition a special phoneme named SIL is included in phoneme sets to represent the Silence features needed by the recognizer tool during the implementation stage.

5.2.3 Feature Vector Extraction

The input of a speech recogniser is the speech waveform. This speech waveform represents all available information. In other words: any set of parameters derived from the input data will contain less or in the best case the same amount of information. The raw input data is however too complex to allow direct modeling, hence the need for a good front-end parameterization which extracts all relevant acoustic information (feature vectors) suitable with the acoustic model designed.

In other words, the preprocessing should remove all non-relevant information such as background noise and characteristics of the recording device, and encode the remaining (relevant) information in a compact set of features that can be handled by the HMM based acoustic models.

Accordingly, feature vector extraction was another important task performed in this experiment. A set of feature files are computed from the speech training data, one for every utterance in the training corpus.

Each speech utterance was transformed into a sequence of features vectors in Mel Frequency Ceptral Co-efficient (MFCC) using a front-end executable provided with the SphinxTrain. Each vector contained 13 cepstral

coefficients with log energy, delta and acceleration coefficients added resulting in 39 dimensional feature vectors. The command used in SphinxTrain is:

perl ./scripts_pl/make_feats.pl –ctl etc/Sidaama_train.fileids

Execution of this command to produce feature vectors needs a file containing the file names of the speech files to be used and a configuration file that determines the parameters for the extraction of feature vectors from the corpus. These files are called 'Sidaama_train.fileids' and 'sphinx_train.cfg' respectively. The MFCC feature files generated were automatically stored in a Sidaama/feats folder.

5.2.4 Dictionary Preparation

It is an already established fact that a dictionary is a valuable resource in the process of building both phoneme and triphone based recognizers. However, manual preparation of a dictionary is time consuming and error prone particularly for building large vocabulary based robust speaker independent recognizer. For technologically advanced languages such as English, commercial and public domain dictionaries are available online in a format suitable for a particular recognizer.

Since there is no such prepared dictionary for the language under study, a phoneme based dictionary is prepared. The same dictionary is used for both context dependent triphone based model and for context dependent monophone based model. This dictionary maps every word in the lexicon to

its corresponding string of phonemes. Example of entries of the phoneme based dictionary is shown in the Table 5.4.

Phoneme based model dictionary	
BUSHA	B U SH A
SHAANA	SH AA N A
YEEDAALA	Y EE D AA L A

Table 5.4: Sample entry in main dictionary file

In addition, the speech recognizer needs a filler dictionary which defines non-speech utterances for the silences between phonemes and words. Basically, this dictionary contains three entry to represent the start of silence <s> in a phoneme or word entry, the middle of utterance silence <sil>, which particularly applies to the phoneme based models and the end of utterance silence </s> in a word or phoneme. In the experimentation all of the silence features were mapped into the same phone SIL which represent silence or background noise. Table 5.5 shows the filler dictionary content.

Filler dictionary content	**Explanation**
<s> SIL	Silence in the beginning of a word/phoneme
<sil> SIL	Silence between words
</s> SIL	Silence at the end of a word /Phoneme

Table 5.5: The contents of filler dictionary

5.2.5 Training HMM model with SphinxTrain

There are several systems that allow building, and training HMM models in order to develop ASR which meets acceptable accuracy, speed and size requirements. One of the most famous and widely used speech recognition tools is the Sphinx system.

In this research SphinxTrain tool which is available with the sphinx system for speech recognition was used to train the HMM models. It is an independent system written in C language common to all sphinx versions, which computes acoustic models from speech records determined for training purpose.

Preparation of HMM takes the feature vectors computed along with their extractions and the phoneme sets that contain the list of phones used to represent the words considered in the study.

Accordingly, in order to create the Sidaama Project environment, Sphinx train tool and the Sidaama project folder were placed in the same directory and on the command prompt of Microsoft windows environment (UNIX or LINUX OS command line terminal such as Cygwin can also be used) the following command is run.

Perl ./SphinxTrain/Scripts_pl/Setup_SphinxTrain.pl –task Sidaama

The HMM creation process is controlled by a file called sphinx_train.cfg. This file is configured to create semi-continuous HMMs model with start and end states topology and three emitting states using single Gaussian density

functions. The states were connected in a left-to-right way, with no skip transitions. A three state topology was chosen because it resulted in better performance for languages that have even longer sounds than the Sidaama phonemes.

Generally in the process of building a given acoustic model the SphinxTrain undergoes the following procedural gradual stages: verification of the input data, HMM training of the context independent (CI) phonemes, HMM training of context dependent (CD) phonemes-triphones, generation of language questions and building the decision tree needed in the tying process of CD phonemes, pruning of the decision tree and tying of similar states, HHM training of the context dependent phonemes with the reduced states number(tied-triphones) and finally training of the tied-triphones models with the gradually augmented number of Gaussians mixtures. In order for the training process to be successfully completed carefully prepared input data must be provided to the training scripts. Pictorial presentation of the sphinx trainer is presented in the figure 5.2.

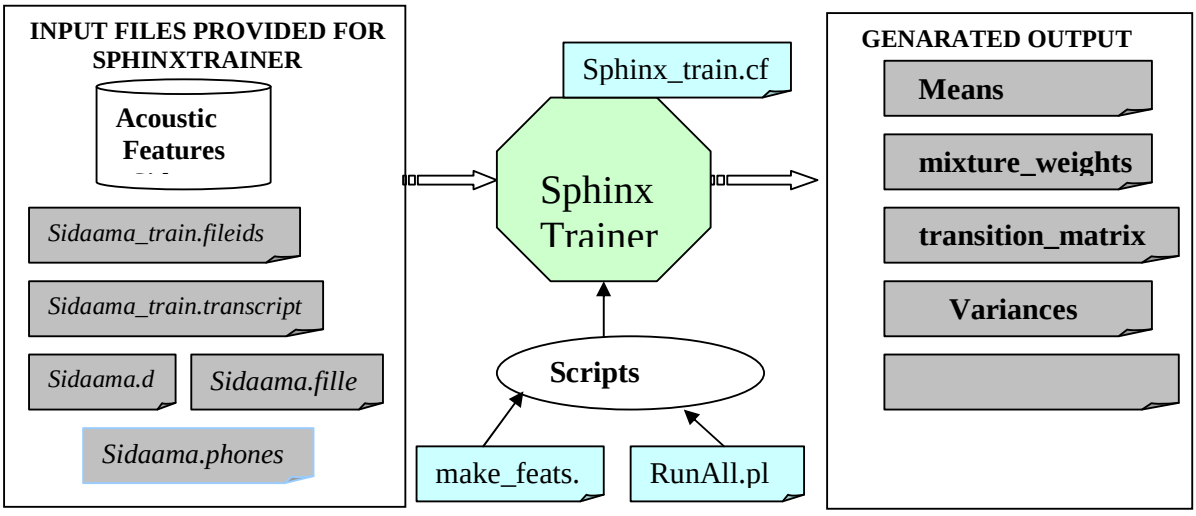

Figure 5. 2: Training procedure of SphinxTrain

The output of SphinxTrain consists of several files which all together represent the acoustic models. These are the CI phonemes with 1 Gaussian mixture, Untied CD phonemes with one Gaussian mixture and tied CD phonemes with multiple Gaussians probability density function (pdf) per states.

The initialization of the models was performed over all training recordings by the so called flat start initialization. Speech files were described by the MFCC parameters and their derivatives. Thus there were 13 cepstral coefficients, 13 delta and 13 delta-delta coefficients for every 20ms frame, computed from 16 bit linear PCM waveforms. During the training the Baum-Welch algorithm was used.

Once all the audio files recorded in a wav format were converted into MFCC format using the make_feats.pl, a number of acoustic models for the Sidaama language were built using the command shown below.

Perl scripts_pl/RunAll.pl

The successful completion of the above command generated both continuous and semi- continuous acoustic models in Sidaama directory called models.

To train the continuous model, parameters of the final 8 Gaussian/ states 3 -state CD –tied acoustic models (HMMs) with 1000 tied states in a directory called model_parameters/Sidaama.Cd_cont_1000_8. The model index files called Sidaama.1000.mdef need for the models are also available in model_architecture folder. This file is used by the system to associate the appropriate HMM parameters with the HMM for each sound unit trained.

5.2.6 The Language Model

Speech recognition can be seen as the task of estimating a maximum a posteriori probability. Given any set of feature files, the system must obtain the most probable sequence of words associated to it. Any sequence of words may be generated by the acoustic models, even some that do not make sense, or that contain grammatical errors, etc.

The purpose of the language model is to make effective use of linguistic constraints when computing the probability of the different possible word sequences. For this task, we used a Unigram based language model, which has been created using the Sphinx Knowledge base language modeling toolkit. It takes a text corpus as an input and produces a trigram or bi-gram models in binary format. These models assume that the probability of any word in the sequence depends only on the previous two words (sample trigram language model is attached in appendix V).

5.2.7 Testing the Recognizers with Sphinx-4 Decoder

The acoustic model developed using SphinxTrain, the language model and the dictionaries are integrated into the Sphinx-4 speech recognizer to perform the actual decoding and testing the performance of prototypes. This needed creating an ANT script in extensible markup language (xml) format that can be compiled using the JDK compiler. A Sample XML configuration script file used to integrate the components of the recognizer is attached in (appendix IV).

Sphinx-4 decoder requires speech test files in raw data format (.raw). The tool switch sound converter is used to convert the test speech files into raw, single channel, 16K sample/second, and 16 bit Big Endean format.

Sphinx-4 also needs transcription batch file which is used as the test reference. It is a main test file where the names, file paths and transcription sentences are given. Sphinx-4 also requires model properties and variable definition files to be properly configured for the decoder to perform. Both files used in the decoding step are presented in appendix I and II.

Once all the necessary configurations are set and inputs provided to the coder, a simple ANT command from the window prompt is used to run the recognition.

As a result the Sphinx-4 decoder produces a recognition hypothesis which is a single best recognition result for each utterance processed. It is a linear word sequence with additional attributes such as respective time, CPU usage, memory requirement and Speed. In addition, the decoder produced a detailed log file that can be used for debugging and computing statistical performance results.

5.2.8 Evaluation Metrics

The Standard evaluation metrics for ASR systems is the Word Error Rate (WER). The word error rate is calculated by comparing the output word string produced by the ASR system called hypothesis with the correct string expected from it called Reference. The comparison is done by calculating the minimum edit distance between the hypothesis and the reference. The

minimum edit distance calculates the minimum substitutions, word insertions and word deletions required to map the hypothesis onto the reference. This minimum edit distance is then converted into WER as follows:

$$\text{Word Error Rates (WER)} = \frac{\textit{Insertions} + \textit{Deletions} + \textit{Substitutions}}{\textit{Total words in Reference Strings}} \text{X100\%}$$

As the expression contains insertions the WER can exceed 100%. Sphinx-4 uses the above method to calculate the trained acoustic model performance. Hence, the recognition results presented in this study used the same principle.

5.3 Experimental Results

This research attempted two independent recognition units, namely, triphone and monophone based units and the following performances were obtained. The performances of the recognizers' were tested based on a corpus of 100 words selected from training set and another 150 unique words not used in training the acoustic models. The recognizers' performance measure undertaken in this research took word accuracy as the main measure of the recognizers' performance.

Accordingly, the recognizer model constructed using context dependent (CD) model using 100 words selected from words used in training set achieved the recognition result shown in table 5.6.

Details	Values
Words	100 utterances
Errors	27 (Sub:11 ,Ins: 0 ,Del: 16)
Accuracy	73.00%
Speed	0.04 X Real time
Total Time	Audio: 191.74s, Processing :7.66s
Memory	Avg. 25.45 MB Max: 35.45 MB

Table 5.6*:* Recognition results for CD model for test words selected from training set

As table 5.6 shows, the prototype recognizer model constructed has correctly classified 73 of the test utterances out of the 100 words achieving 73% accuracy for triphone based model. This result also indicated that out of the total 27 wrongly classified words 11 of them were reported as substitution, 16 as deletion, and non as insertion errors.

Similarly, the context dependent model constructed was tested using another 150 test words which are not used in training the acoustic model and the performance obtained is shown in table 5.7.

Details	Values
Words	150 utterances
Errors	63 (Sub: 37 ,Ins: 0 ,Del: 26)
Accuracy	58.00%
Speed	0.06 X Real time
Total Time	Audio: 191.74s, Processing :10.67s
Memory	Avg. 35.47 MB Max: 57.45 MB

Table 5.7. Recognition result of CD model for test words not in the training set

Most of the errors made during the recognition phase of the context dependent model appear to be deletion and substitution errors. Substitution errors are errors that happen when the utterances are wrongly transcribed. For instance, the HMM model transcribed the speech 'SHIIMA' for 'SHAANA, and 'SHAANA' as ' SHAANA', hence replacing the word 'SHIIMA by 'SHAANA'. This usually happens because of the phoneme similarity between the two words. Such kind of recognition error can be minimized by training the recognizer with more data so that the transition probability for the phoneme can be adjusted and the recognizer easily identifies the two words. Additionally, better feature extraction techniques can be used to avoid or minimize the level of confusion of the recognizer by producing unique feature vectors for each of the utterances used.

Insertion errors however introduce new words into the recognition output. For example the word 'JAWATA' was transcribed as ' JAWATA, KINCHO'. This shows that the word 'KINCHO' was wrongly inserted. One reason that can be attributed to this effect is the poor quality of the audio data. Even though quit room was selected during recording the speech data, external noises such as environmental noise and noise coming form the recording computer could significantly affected the quality of the test data. Where as, deletion errors are those errors that remove words from the recognition sentences or outputs.

The second prototype model built in this study is the context independent recognizer model that used monophones as recognition units. For testing the recognition accuracy of context independent model the same amount of test

data was used as that of the context dependent model. The recognition performance result obtained for monophone based context independent model using 100 tests sets selected from the training set is presented in table 5.8 as follows.

Details	**Values**
Words	100 utterances
Errors	35 (Sub:17, Ins:4, Del: 14)
Accuracy	65.00%
Speed	0.21 X Real time
Time	Audio: 191.74s, Processing:80.64s
Memory	Avg: 64.16 Mb Max: 128.38 MB

Table 5.8*:* Recognition result of CI model for test words in training set

As shown in Table 5.8, the context independent recognizer model achieved a recognition accuracy of 65 % for tests words used from the training set. Most of the errors displayed in this model were deletion and substitution errors.

On the other hand the recognition performance result obtained for monophone based context independent model using 150 tests sets which are not used in training the acoustic model is presented in table 5.9 as follows.

Details	Values
Words	150 utterances
Errors	78 (Sub:42, Ins:2, Del: 34)
Accuracy	48.00%
Speed	0.21 X Real time
Time	Audio: 191.74s, Processing:90.64s
Memory	Avg: 64.16 Mb Max: 128.38 MB

Table 5.9. Recognition results of CI model for words not in training set

Similarly the recognizer resulted higher recognition time and more memory consumptions as compared to the context dependent model. Considering the recognition accuracy, recognition time and memory consumptions, it can be concluded that the context independent monophone based model showed more errors than the triphone based recognizer model.

As can be observed from the results of the experiments both models achieved promising results as this was the first attempt in trying to build speaker dependent recognizer model for Sidaama language.

Although in both recognition units the results obtained were not close to the standard accuracy level achieved in building the desired speaker dependent recogniser system, taking in to account the data used for training the models, the parameters used to train the models, the recording environment and the available time it is a highly encouraging result. However, by using more training data and changing the model parameters, results obtained in this research may significantly be improved in the future.

Comparing the results obtained from testing the two recognizer models built in this research, the context dependent triphone based model achieved better recognition accuracy. As a result, it is selected to be a good candidate for building robust recognition systems for the Sidaama language.

Although this experimental work was challenged by lack of digitally recorded speech corpus and exhaustive information about the rules of the language, the constructed prototype Sidaama speech recognizer model revealed encouraging outcomes.

In conclusion, the findings of this research indicated that triphone based recognition units are appropriate for building recognizer for the Sidaama language as compared to monophone based model. The most errors encountered in testing the two recognizers were substitution and deletion errors. These errors can be minimized by extensively training the models with large corpus. Regarding decoding speed and memory usage the context dependent triphone based model also showed a good result than that of the context independent monophone based model.

CHAPTER SIX

CONCLUSIONS AND ECOMMENDATIONS

6.1 Conclusions

Speech recognition systems have been applicable in wide areas as various speech recognition methodologies, techniques and tools have been developed and implemented to generate a natural and intelligible speech. Moreover, these technologies have also been advanced to incorporate speech recognition systems in different computing devices such as mobile phones, ATM machines etc.

In this research, first an attempt was made to explore the possibility of developing prototype speech recognition for Sidaama language using Hidden Markov Model. In order to come up with a working prototype model for the language extensive study was conducted on the language to understand and come up with the language features needed to build the model. However, only some features of the language were studied since it requires a lot of time and deep knowledge to exhaustively study the features and characteristics of the language.

Additionally, the components as well as techniques used in the HMM based speech recognition design were studied and analyzed to identify those components that are dependent on the characteristics of the language. Besides the most commonly used speech recognition tools were critically reviewed and as a result the most widely used speech recognizer tool called the Sphinx Systems was used. SphinxTrain which is one the sphinx systems

components was used to build the acoustic models where as for testing the recognizers' performance a java based font-end decoder called Sphinx-4 was employed.

This research attempted to build context dependent triphone based speech recognizer as well as context independent monophone based isolated word speech recognizer models for Sidaama. A total of 450 speech utterances were recorded using 450nunique words from a single speaker. Out of the total datasets recorded, 300 speech utterances were used for training the acoustic models. Whereas, the remaining 150 words of speech not included in training the recognizer models were used for testing the accuracy of the recognizer models. In addition 100 words of speech were randomly selected from the training set in order to test the recognizer accuracy. The same proportion and amount of training and test data were used in building the two recognizer models (triphone based and monophone based models).

The constructed models were tested with the same amount of test data and achieved promising results. The context dependent triphone based model achieved recognition accuracy of 73% for test words selected from training set and 68% accuracy for test words not included in training set . Whereas the context independent monophone based model achieved a recognition accuracy of 65% for test words selected from the training set and 48% accuracy for tests words not included in training set. Comparing the results obtained it can be concluded that the triphone based model outperformed the monophone based context independent model for the language. The

triphone based model also showed a better processing time and memory usage than the monophone based context independent model.

In general the results obtained from this study were encouraging and the outcomes obtained can be taken as a good start for further research in the area of speech recognition for Sidaama language.

6.2 Recommendations

Based on the findings of the study and to improve the recognizer models developed for Sidaama language we recommend the following future considerations.

- In order to improve the performance of the context independent recognizer model large amount of training data should be used to better estimate the probability parameters used in HMM
- this study was disadvantaged by unavailability of recorded speech corpus for the language. Hence, future attempts should be made to prepare standard speech corpus for the language and further large vocabulary based experiments should be conducted
- this study used monophones and triphones as recognition units to get better results other recognition units such as syllables should also be tested
- this study employed Statistical based HMM approaches in building the system, however neural network based or Hybrid approaches should also be tested to study the performance of a recognizer model.

Reference:

Abebe Gebre-Tsadik (1985) Derived Nominals in Sidamo, Undergraduate Thesis. Addis Ababa University.

Anusuya, M.A., and Katti,S.K.(2009). Speech Recognition by a Machine: International Journal of Computer Science and Information Security, Vol.6, No.3.

Anbessa Teferra(2000).A grammar of Sidaama, Doctoral dissertation. Jerusalem: The Hebrew University.

Ashenafi Demisse.(2009).A Speech Recognition System for Afaan Oromo, Master's Thesis, Addis Ababa University, Addis Ababa.

Cole, R.A , Mariani, J.,Uszkoreit, H, Zaenen, A. and Zue, V.(1995): Survey of the State of the Art in Human Language Technology, Center for Spoken Language Understanding CSLU, Carnegie Mellon University, Pittsburgh, PA.

Comez, M.A.(2003). Large Vocabulary Continuous Speech Recognition for Turkish Language using HTK. Master Thesis. School of Natural and Applied Science of Middle East Technical University, Istanbul**.**

Cook, S., (2003). Speech Recognition-How to. Available at: http://www.faqs.org/docs/Linux-HOWTO/Speech-Recognition HOWTO.html.

Forsberg M. (2003). Why Speech Recognition is Difficult? Department of Computing Science, Chalmers University of Technology, Gothenburg, Sweden.

Furui, S.(2005). " 50 years of Progress in Speaker Recognition Research".ECTI Transaction on Computer and Information Technology.Vol.1, No.2.

Furui, S.(2001).Digital Speech Processing, Synthesis, and Recognition-Second Edition New York –USA, Marcel Dekker Inc.

Furui, S. (1995). Toward the ultimate synthesis/recognition systems. In Proceedings of the National Academy of Science. Vol. 92, pp. 10040-10045.

Hafte Abera.(2009). Hidden Markov Model Based Large Vocabulary, Speaker Independent Continuous Tigrigna Speech Recognition. Master's Thesis, Addis Ababa University, Addis Ababa.

Hansen, J.C.(2003). Modulation Based Parameters for Automatic Speech Recognition. Master's Thesis, Department of Electrical Engineering, University of Rhode Island, USA.

Harmensky , H (1990)."Perceptual Linear Predictive (PLP) Analysis of Speech" Journal of the Acoustical Society of America, Vol.87,No. 4, pp 1738-1752.

Census,(2008).The Summary and Statistical Report of the 2007 Population and Housing Census of Ethiopia: Analytical Report at National Level, Addis Ababa, Ethiopia.

Helin, D.(1999). Statistical language models for large vocabulary Turkish Speech recognition. Master Thesis. Bogazici University.

Holmes, J., Holmes, W., (2003).Speech Synthesis and Recognition. Taylor and Francies New Fetter Lane, London ECAP 4EE.

Holmes, J. N. (2001). Speech Synthesis and Recognition. London: Taylor and Francis.

Honda, M.(2003).Human Speech Production Mechanism. NNT Technical Review, Vol.1, No.2.

Hudson, Grover (2005). Highland East Cushitic languages,Encyclopedia of Language and Linguistics, 2nd ed., Keith Brown, ed., 294-298. Elsevier: Oxford.

Jackson, M. (2005). Automatic Speech Recognition: Human Computer Interface for Kinyarwanda Language. Master Thesis. Makerene University.

Jelinek ,F.(1997). "Statistical Methods for Speech Recognition", The MIT Press.

Juang, B. H. and Lawrence R. Rabiner (2004). "Automatic Speech Recognition A Brief History of the Technology Development". Elsevier Encyclopedia of Language and Linguistics, 2nd ed.

Juang, B.H and Rabiner,R (2004).Automatic Speech Recognition-A brief History of Technology Development. Georgia Institute of Technology, Atlanta.

Juang ,B.H and Rabiner,R(1991). Technometrics, Vol. 33, No.3, pp. 251-272

Jurafsky, D. and Martin, J.(2000). Speech and Language Processing, Upper Saddle River, NJ, USA: Prentice-Hall

Jurafsky, D. and Martin James. H (,1999). Speech and Language Processing – An Introduction to natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall Press.

Kim, Woosung (2004). Language Model Adaption for Automatic Speech Recognition and Statistical Machine Translation, PhD thesis, Johns Hopkins University, Baltimore, ,Maryland

Kawachi, Kazuhiro (2007) A grammar of Sidaama, a Cushitic language Ethiopia. PhD thesis. State University of New York at Buffalo.

Kinfe, T.,(2001).Sub-word Based Amharic Word Recognition: An Experiment Using Hidden Markov Model (HMM). Master's Thesis, Addis Ababa University, Addis Ababa.

Lamere, P., Kwok P., Walker, W., Gouvea, E., Singh, R. Raj, B., Wolf, P(2003). Design of CMU Sphinx-4 Decoder. Sun Microsystems Laboratories, Carnegie Mellon University, Mitsubishi Electronic Research Labratories, USA.

Laver. J. (1990): Principles of Phonetics. Cambridge University Press, 1994.

Levinson, S.E, Rabiner L.R., and Sondhi M.M (1988): An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. The Bell System Technical Journal, Vol.62, No.4.

Lee, C.H, and Juang, B.H (1996). A Survey on Automatic Speech Recognition with an Illustrative Example on Continuous Speech Recognition of Mandarin. Computational Linguistics and Chinese language Processing. Vol.1, No.1.

Lee. K-F, (1990). Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. IEEE Trans. Acoust. Speech Signal Processing, 38(4).

Lee, K.,(1989). Automatic Recognition. Boston: Kluwer Academic publishers.

Magdi, A.M. and Gader, P.(2000).Generalized Hidden Markov Models-PartI: Theoretical Frameworks. IEEE Transaction on Fuzzy Systems, Vol.8, No.1.

Markowitz, J.,(1996).Using Speech Recognition.Upper Saddle River, New Jersey:Prentice Hall, Inc.

Makhoul, J., and Schwartz, R. (1995). State of the art in continuous speech recognition. In Proceedings of the National Academy of Science. Vol. 92, pp. 9956-9963.

Martens, J.P.(2000).Continuous Speech Recognition over the Telephone. Electronics and Information Systems, Ghent University, Belgium.

Martha Y.,(2003).Application of Amharic speech recognition system to command and control computer: an Experiment with Microsoft Word: Master's Thesis, Addis Ababa University, Addis Ababa

Ming, T., C.(2007) Malay continuous speech recognition using continuous density hidden Markov model. Masters thesis, Universiti Teknologi , Malaysia.

Olsen, S.2003. "Fundamentals of Lingustics Analysis", Available at: http://www2.huberlin.de/angl/oslen-analysis.pdf.

Philip, G and E. S. Young(1987). "Man-machine Interaction by voice: Developments in Speech Technology" Journal of Information Science vol. 13. pp. 3-14.

Podder, S.K. (1997).Segment-based Stochastic Modeling for Speech Recognition. PhD Thesis. Department of Electrical and Electronic Engineering , Ehime University, Matsuyama 790-77, Japan.

Plannerer, B, (2005), Introduction to Speech recognition- Edition 1.1, Munich-Germany, IEEE publication.

Rabiner & Juang(1993).Fundamentals of Speech Recognition .Prentice Hall.

Rabiner,L.R(1989)."A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". Proceedings of the IEEE. pp 257-285.

Radman, R.(1999).Computer Speech Technology. Norwood: Artech House, Inc.

Reiderer, K.(1999)."Large Vocabulary Continuous Speech Recognition"Helsinki University of Technology: Laboratory of Computational Engineering.

Roach, P." A little encyclopedia of phonetics", University of Reading,UK,2002.

Shaefer, R.W.(1995). Scientific bases of human machine communication. In proceedings of the National Academy of Science. Vol.92,pp. 9914-9920

Satori, H., Harmi, M.,and Chenfour, N(2003). Introduction to Arabic Speech Recognition Using CMU Sphinx Systems.

Samsudin, N., Tiun, S., and Kong, T.,(2003). A Simple Malay Speech Synthesizer using Syllable Concatenation Approach. Available at: http://www.cs.bham.ac.uk.

Schiel,F. and Draxler, C. (2004).The production of speech corpora, BITS Projekt-Account. Available at: www.phonetic.unimuenchen.de/BITS/TP1/CookBook/Tp1.html

Spaans, M.A.(2004).On Developing Acoustic Models Using HTK. Master's Thesis. Delft University of Technology.

Solomon T.,Martha,Y. and Menzel, W. (2008) "Amharic Speech Recognition: Past, Present, and Future". University of Hamburg, Germany.

Solomon Berhanu (2001): Isolated Amharic Consonant –Vowel (CV) Syllable Recognition. An experiment using the Hidden Markov Model. Master's Thesis. Addis Ababa University.

Tebelskis,J(1995). Speech Recognition using Neural Networks. School of computer Science. Carnegie Mellon University. Pittsburgh, Pennsylvania.

Ursin, M.,(2002).Triphone clustering in Finnish continuous speech recognition , Master's Thesis ,Helsinki University of Technology, Department of Computer Science, Helsinki.

Walker, W.,Lamere,P.,Kwok,P.,Raj,B.,Singh,R.,and Woelfel,J(2004).Sphinx4. A Flexible Open Source Framework for Speech Recognition.-Technical Report, Sun Microsystems Inc.

Wiggers , P.(2001). Hidden Markov Models for Automatic Speech Recognition and their Multimedia Applications. Delft University of Technology, Netherlands.

Wong, P. H. W. et al. (1998). Reducing Computational Complexity of Dynamic Time Warping Based Isolated Word Recognition with Time Scale Modification. Proceedings of 1998 Fourth International Conference on Signal Processing Proceedings. vol.1: 722 -725.

Zhang, M.,(2001).Overview of Speech Recognition and Related Machine Learning Techniques-Technical Reports, School of Mathematical and Computer Sciences, Victoria University of Willington.

Zegaye, S.,(2003).Hidden Markov Model Based Large Vocabulary, Speaker Independent, Continuous Amharic Speech Recognition. Master's Thesis, - Faculty of Informatics, Addis Ababa University.

Zue, V and R. Cole. (1995) "Spoken Language Input: Overview." Survey of the State of the Art in Human Language Technology. Eds. Cole, R.A et al., Oregon Graduate Institute.

Zue V, Cole .R, & Ward, W.(2001). Speech Recognition – Defining the problem, MIT Laboratory for Computer Science, Cambridge, Massachusetts, Oregon Graduate Institute of Science & Technology, Portland, Oregon, and Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Zoric, G. (2005). Automatic Lip Synchronization by Speech Signal Analysis. Master Thesis, Department of Telecommunications, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia.

Appendices:

Appendix I: Variables definition file used in Sphinx-4 (variable. def)

```
set exptname = Sidaama
set vector_length = 13
set dictionary = $base_dir/lists/Sidaama.dic
set fillerdict = $base_dir/lists/Sidaama.filler
set statesperhmm = 3
set skipstate = no
set gaussiansperstate = 8
set feature = 1s_c_d_dd
set n_tied_states = 1000
set cmn = current
set varnorm = no
```

Appendix II: Model properties File used in Sphinx-4 (Model.Props)

```
description = Sidaama Acoustic Models
modelClass =
edu.cmu.sphinx.model.acoustic.Sidaama_8gau_13dCep_16k_40mel_130Hz_6800Hz.Model
modelLoader =
sphinx.model.acoustic.Sidaama_8gau_13dCep_16k_40mel_130Hz_6800Hz.ModelLoader
dataLocation = cd_continuous_8gau
modelDefinition = etc/Sidaama_8gau_13dCep_16k_40mel_130Hz_6800Hz.1000.mdef
isBinary = true
featureType = 1s_c_d_dd
vectorLength = 39
sparseForm = false
numberFftPoints = 512
numberFilters = 40
gaussians = 8
minimumFrequency = 130
maximumFrequency = 6800
sampleRate = 16000
```

Appendix III: Sample words used to build the recognizer prototype Model

AGURA	ANNOO'MA	CAABBICHCHO	IIBBABO	OFOLLAANCHO
AJIIFFATA	ARGIRA	CAACCAAWA	IIBBADO	OLANSHO
AKAAKO	ARRAWO	CAAFA	IILLISHA	OLLAA
AKKA'MA	ARSHAMMO	CAALE	IJAADHA	ONTAA
ALBISA	ASSAADDO	CAAQA	ILAMME	ONTE
ALEECHO	ASSAANCHO	CAAQO	ILLISAMA	OOFAANCHO
ALEEGE	ATOOTE	CANCA'NA	IMAANA	OOMBARRA
AMAXXA	AWAAJJA	CEE'MA	IMAANSIISA	QAROOYYE
AMBOOMA	AWUUWA	CEEHAMA	IRREESSA	QASAMA
AMEESSA	AYIDDE	CHENCHE	ITI	QATIFATA
AMMA'NA	BA'LA	CHIGGINYE	ITISAANCHO	QAWAAXXO
ANGA	BA'LATTO	CHUUCO	IWIPHPHA	QAWWE
GAADISAANCHO	BAAJEETE	CII'MA	IYYAHO	QAXXADHA
GAADISIISA	BAANKE	CIMBA	IYYAQQA	QEELLE
GAADO	BAAQULCHO	CIWAA	JAALOO'MA	QEEQQA
GAAGGO	BAATISIISA	CODHO	JAAMMISA	QINAAWO
GAAGIISSA	BAAXXA	CORQA	JAANJIWEELO	QIXXEESSAANCHO
GAALLAANO	BADHEESSA	CUBBO	JADDEESSA	RAGAMA
GAALSIISAANCHO	BALALLE	CUQQAALAMA	JAJJOO'MA	RAKKAMA
GAANA	BALATTE	CUWE	JARANA	RIBBO
GAANGAANA	BALLOO'MA	DAAFURA	JAWAATA	RIICO
GAARE	BAQQAALLA	DAAHI	JAWWITTA	RODIIWEELA
GABBARO	BARAARA	DAALASA	JIBBAATA	RODOO
GABBEESSA	BARBARCHO	DAANIICHCHO	JIGEESSA	ROOSIISAANCHO
GABBISO	BARKIRA	DADDA'LA	JISSA	ROPHILA
GACCO	BARMEELE	DADHAMA	JOONGIMMA	ROSIISA
GADAANCHO	BARNEEXA	DAFAANCHO	KAATAANCHO	ROSO
GADADIISA	BARRA	DAHEESSA	KARSIIRA	ROYI'RA
GALAAFFATA	BAWWE	DANANCHO	KEWELCHO	SAGADA
GALAWA	BAXILLE	DARAAFFISA	KIBBA	SAJJOO
GALFATA	BAYI'RA	DARAARA	KIIRO	SAMBATA
GALLAJJO	BAYIRA	DAYYAASA	KIPHO	SARPHAPHO
GALTIWEELA	BEEDDAKKO	DEE'NIYAA	KONNE	SASSOO
GAMACCA	BEEQQO	DII'RE	LAAFA	SEEJJA
GAMAXA	BERO	DIILALLA	LAAGUXA	SEEKKISIISA
GAMMEENA	BIILLE	DIINA	LAANCHA	TIRSIISA
GANANSHO	BILBILA	DINNICHCHA	LAAYYAAWA	TIYYI'RA
GANCARA	BISKLEETE	DINYNYE	LALLAWO	TIYYO
GANYNYA	BOHAMA	DIRRO	LAPHPHISE	TUNSIISA
GARGADHA	BOKKA	DO'NOOTA	LEELLAMA	TUTUMME
GARROTE	BOOLAQA	DOGGANCHO	LEKKA	UGAAXA
GASHA	BOOLO	DOOYYE	LEXXA	UGGE
GASHOOTE	BOOTTA	DOQIDOQQE	LIIQAAWA	URGEESSA
GASHSHAANCHO	BOOWE	DUDUWO	LIMIXEESSA	WAACCITO
GEEDIWEELLA	HAADA	DURRIISA	LIXA	WAADANCHO
GEEGIISHSHA	HAANJA	EELA	LOOSIISAANCHO	WAAJJA
GEERAARA	HAAQE	EERA	MAAXANSHO	WAASA
GEERCHO	HAAWIISSA	EESSAANCHO	MACCA	XAADISA
GEERSIISA	HACULA	EEYYA	MAGAALA	XAARO
GEESHSHA	HAGGO	EGEMMA	MALAATISA	XAGICHCHO
GEEXANO	HAGIIDHA	EGENNA	MALLAANCHO	XAWO
GELEELCHA	HAINI	ELEELLE	MANAXXIRE	XEENA
GELEGE'LA	HAJAJO	ELLA	MANGISTE	XELLEQQO
GELFAAWA	HALALLA	EREEDA	MAQASE	XIINXALLA
GEYYO	HALCHO	EREERO	MAXAAFICHCHO	YAADDAAWA
			NAAXXA	YAANCHO

Appendix IV. Sample configuration file used in Sphinx-4 (config.xml)

```
<?xml version="1.0" encoding="UTF-8"?>
<!--
   Sphinx-4 Configuration file
-->
<!-- ******************************************************** -->
    <!-- The Dictionary configuration                         -->
    <!-- ******************************************************** -->
    <component name="dictionary"
               type="edu.cmu.sphinx.linguist.dictionary.FastDictionary">
        <property name="dictionaryPath"

value="resource:/edu.cmu.sphinx.model.acoustic.Sidaama_8gau_13dCep_16k_40me
l_130Hz_6800Hz.Model!/edu/cmu/sphinx/model/acoustic/Sidaama_8gau_13dCep_16k
_40mel_130Hz_6800Hz/dict/Sidaama.dic"/>
        <property name="fillerPath"

value="resource:/edu.cmu.sphinx.model.acoustic.Sidaama_8gau_13dCep_16k_40me
l_130Hz_6800Hz.Model!/edu/cmu/sphinx/model/acoustic/Sidaama_8gau_13dCep_16k
_40mel_130Hz_6800Hz/dict/Sidaama.filler"/>
        <property name="addSilEndingPronunciation" value="false"/>
        <property name="allowMissingWords" value="true"/>
        <property name="unitManager" value="unitManager"/>
    </component>

    <!-- ******************************************************** -->
    <!-- The Language Model configuration                       -->
    <!-- ******************************************************** -->
    <component name="ngramLanguageModel"

type="edu.cmu.sphinx.linguist.language.ngram.SimpleNGramModel">
        <property name="location" value="Sidaama.flat_unigram.lm"/>
        <property name="logMath" value="logMath"/>
        <property name="dictionary" value="dictionary"/>
        <property name="maxDepth" value="1"/>
        <property name="unigramWeight" value=".7"/>
    </component>

    <!-- ******************************************************** -->
    <!-- The acoustic model configuration                       -->
    <!-- ******************************************************** -->
    <component name="Sidaama"

type="edu.cmu.sphinx.model.acoustic.Sidaama_8gau_13dCep_16k_40mel_130Hz_680
0Hz.Model">
        <property name="loader" value="sphinx3Loader"/>
        <property name="unitManager" value="unitManager"/>
    </component>

    <component name="sphinx3Loader"

type="edu.cmu.sphinx.model.acoustic.Sidaama_8gau_13dCep_16k_40mel_130Hz_680
0Hz.ModelLoader">
        <property name="logMath" value="logMath"/>
        <property name="unitManager" value="unitManager"/>
    </component>
```

Appendix V: Sample Language Model for Sidaama

Language model created by QuickLM on Mon Jul 5 11:57:11 EDT 2010
Copyright (c) 1996-2010 Carnegie Mellon University and Alexander I. Rudnicky

The model is in standard ARPA format, designed by Doug Paul while he was at MITRE.

The (fixed) discount mass is 0.5. The backoffs are computed using the ratio method.
This model based on a corpus of 457 sentences and 460 words

\data\
ngram 1=460
ngram 2=915
ngram 3=459

\1-grams:

-0.7788 </s> -0.3010
-0.7788 <s> -0.2220
-3.1377 A -0.2220
-3.4387 AADA -0.2220
-3.4387 AADDO -0.2220
-3.4387 AAGANKA -0.2220
-3.4387 AAGINTAAWA -0.2220
-3.4387 AAKKO -0.2220
-3.4387 ABBISIISAANCHO -0.2220
-3.4387 AFAALEESA -0.2220
-3.4387 AFANYNYA -0.2220
-3.4387 AFOO -0.2220
-3.4387 AFUDUUDA -0.2220
-3.4387 AGAAGULE -0.2220
-3.4387 AGADHA -0.3007
-3.4387 AGARSIISI'RA -0.2220
-3.4387 AGGAXXA -0.2220
-3.4387 AGURA -0.2220
-3.4387 AJIIFFATA -0.2220
-3.4387 AKAAKO -0.2220
-3.4387 AKKA'MA -0.2220
-3.4387 ALBISA -0.2220
-3.4387 EEYYA -0.2220
-3.4387 EGEMMA -0.2220
-3.4387 EGENNA -0.2220
-3.4387 ELEELLE -0.2220
-3.4387 ELLA -0.2220
-3.4387 EREEDA -0.2220
-3.4387 EREERO -0.2220
-3.4387 FAYYA -0.2220
-3.4387 FAYYIMMA -0.2220
-3.4387 FIDALE -0.2220
-3.4387 FIIGANCHO -0.2220
-3.4387 FIIXA -0.2220
-3.4387 FIIXOONSA -0.2220
-3.4387 FILA -0.2220
-3.4387 GADAANCHO -0.2220
-3.4387 GADADIISA -0.2220
-3.4387 GALAAFFATA -0.2220
-3.4387 GALAWA -0.2220
-3.4387 GALFATA -0.2220
-3.4387 GALLAJJO -0.2220
-3.4387 GALTIWEELA -0.2220
-3.4387 GAMACCA -0.2220
-3.4387 GAMAXA -0.2220
-3.4387 GAMMEENA -0.2220
-3.4387 HAADA -0.2220
-3.4387 HAANJA -0.2220

-3.4387 ALEECHO -0.2220	-3.4387 HAAQE -0.2220
-3.4387 ALEEGE -0.2220	-3.4387 HAAWIISSA -0.2220
-3.4387 AMAXXA -0.3007	-3.4387 HACULA -0.2220
-3.4387 AMBOOMA -0.2220	-3.4387 HAGGO -0.2220
-3.4387 AMEESSA -0.2220	-3.4387 HAGIIDHA -0.2220
-3.4387 AMMA'NA -0.2220	-3.4387 HAINI -0.2220
3.4387 BAATISIISA -0.2220	-3.4387 HAJAJO -0.2220
-3.4387 BAAXXA -0.2220	-3.4387 IILLISHA -0.2220
-3.4387 BADHEESSA -0.2220	-3.4387 IJAADHA -0.2220
-3.4387 BALALLE -0.2220	-3.4387 ILAMME -0.2220
-3.4387 BALATTE -0.2220	-3.4387 ILLISAMA -0.2220
-3.4387 BALLOO'MA -0.2220	-3.4387 IMAANA -0.2220
-3.4387 BAQQAALLA -0.2220	-3.4387 IMAANSIISA -0.2220
-3.4387 BARAARA -0.2220	-3.4387 IRREESSA -0.2220
-3.4387 BARBARCHO -0.2220	-3.4387 JARANA -0.2220
-3.4387 BARKIRA -0.2220	-3.4387 JAWAATA -0.2220
-3.4387 BARMEELE -0.2220	-3.4387 JAWWITTA -0.2220
-3.4387 BARNEEXA -0.2220	-3.4387 JIBBAATA -0.2220
-3.4387 BARRA -0.2220	-3.4387 JIGEESSA -0.2220
-3.4387 BAWWE -0.2220	-3.4387 JISSA -0.2220
-3.4387 BAXILLE -0.2220	-3.4387 JOONGIMMA -0.2220
-3.4387 CANCA'NA -0.2220	-3.4387 KAATAANCHO -0.2220
-3.4387 CEE'MA -0.2220	-3.4387 KARSIIRA -0.2220
-3.4387 CEEHAMA -0.2220	-3.4387 KEWELCHO -0.2220
-3.4387 CHENCHE -0.2220	-3.4387 KIBBA -0.2220
-3.4387 CHIGGINYE -0.2220	-3.4387 LAAGUXA -0.2220
-3.4387 CHUUCO -0.2220	-3.4387 LAANCHA -0.2220
-3.4387 DADHAMA -0.2220	-3.4387 LAAYYAAWA -0.2220
-3.4387 DAFAANCHO -0.2220	-3.4387 LALLAWO -0.2220
-3.4387 DAHEESSA -0.2220	-3.4387 LAPHPHISE -0.2220
-3.4387 DANANCHO -0.2220	-3.4387 XAGICHCHO -0.2220
-3.4387 DARAAFFISA -0.2220	-3.4387 XAWO -0.2220
-3.4387 DARAARA -0.2220	-3.4387 XEENA -0.2220

DECLARATION

THIS THESIS IS MY ORIGINAL WORK AND HAS NOT BEEN SUBMITTED AS A PARTIAL REQUIREMENT FOR A DEGREE IN ANY UNIVERSITY, THAT ALL SOURCES OF MATERIAL USED FOR THE THESIS HAVE BEEN DULY ACKNOWLEDGED.

_____

ABDELLA KEMAL MOHAMMED

JULY, 2010

THE THESIS HAS BEEN SUBMITTED FOR EXAMINATION WITH MY APPROVAL AS UNIVERSITY ADVISOR.

_____

DEREJE TEFERI (Ph.D.)