ADDIS ABABA UNIVERSITY

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

A CONTINUOUS, SPEAKER INDEPENDENT SPEECH RECOGNIZER FOR AFAAN OROMOO

BY

KASSAHUN GELANA MICHO

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfilment of the Requirement for the Degree of Masters of Science in Information Science.

JULY, 2010

ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

A CONTINUOUS, SPEAKER INDEPENDENT SPEECH RECOGNIZER FOR AFAAN OROMOO

BY

KASSAHUN GELANA MICHO

Signature of Board of Examiner for Approval

_____	_____
Chairman, Department Graduate committee	Signature
_____	_____
Advisor	Signature
_____	_____
Examiner	Signature

Acknowledgement

Above all, I would like to praise my almighty God, who enabled me succeed in my work. I am also very much grateful to my advisor Ato Mulu G/Egziabher for his constructive advice and critical comments from the beginning of this research work to the end of it. His valuable advice and comments helped me a great deal in shaping the paper.

My families and friends, Mr. Akalu Z. And Dr.Daniel B., should also deserve my deep honor for their valuable encouragements and remarkable supports provided to me when conducting the research work.

Equally, I wish to give my heartfelt gratitude to my friends and colleagues in AAU Alemayehu Tilahun, Abdella Kemal, Habtamu Fanta and all my classmates who helped me a lot especially in commenting on the work and in sharing books, information and other documents from the beginning to the end of this research work.

Finally I would like to say thank you for Adanech Huluka for your contribution throughout my study.

Dedication

To my Family

Abstract

The ultimate goal of any automatic speech recognition is towards developing a model that converts speech utterance to texts words. Therefore, a continuous, speaker independent Afaan Oromoo speech recogniser's experiment is performed having similar objective of transforming Afaan Oromoo continuous speech in to its text word formats for continuous Afaan Oromoo speaker independent speech utterances using HMM and sphinx system (sphinx train for training and Sphinx4 for decoding). Therefore, this research tries to develop prototype for a continuous, speaker independent Afaan Oromoo speech recognizer so as to check possibility and suitability of the tools and techniques selected from the various literatures.

A continuous, speaker independent Afaan Oromoo speech recognizer is developed in this research work for 70 selected Afaan Oromoo long words, phrase and simple sentence uttered by 30 selected peoples from different age group and sex constituting of 2100 utterances. Accordingly, the data collected was divided in the 2/3 by 1/3 for training and testing respectively. Furthermore, various pre-processing and other activities were performed including building the acoustic and language models among others which might greatly affect significantly the performance of the recognizer. These Afaan Oromoo selected words, phrases and simple sentences are selected in consultation of the domain experts.

For this research performance evaluation is performed using test data sets and the recognizer performance is found to be 68.514% with sentence accuracy of 28% for continuous Afaan Oromoo speech and a phoneme based trigram performance of 89.459% with sentence accuracy of 42% achieved.

According to the finding of this research, the performance gained for Afaan Oromoo language is highly promising and as the language is becoming one of the most spoken language of the country, it will have tantamount for latter full deployment of the recognizer in the language.

Keywords: Speech recognition, Afaan Oromoo, sphinx, Hidden Markov Model

Table of contents

Table of Contents **Pages**

Lists of Acronyms

AI	Artificial Intelligence
AM	Acoustic Model
AP	Acoustic Phonetic
ASR	Automatic Speech Recognition
BFCC	Bark Frequency Cepstral Coefficients
CAOSR	Continuous Afaan Oromoo Speech Recognition
CD	Context Dependent
CI	Context Independent
CMU	Carnegie Mellon University
DCT	Discrete Cosine Transformation
DFT	Discrete Fourier Transform
DTW	Dynamic Time Warping
HTK	Hidden Markov Model Toolkit
HMM	Hidden Markov Model
IPA	International Phonetic Alphabet
ISR	Isolated Word Speech Recognition
IWR	Isolated Word Recognition
KHz	Kilo Hertz
KS	Knowledge Source
LM	Language Model
LPC	Linear predictive coding
LVQ	Learning Vector Quantization
MERL	Mitsubishi Electric Research Laboratory
MFCC	Mel Frequency Cepstral Coefficients
MLP	Multi Layer Perception
PLP	Perceptual Linear Prediction
PDF	Probability Density Function
RSI	repetitive strain injuries
SRM	Structural Risk Minimization
SVM	Support vector machine
WER	Word Error Rate
VQ	Vector Quantization

Lists of Figure

Lists of Tables

CHAPTER ONE

1.1. Introduction

Speech is a natural form of communication for human beings, and computers with the ability to understand speech and speak with a 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, through the evolution of speech synthesis and speech recognition technologies (Honda, 1999). Therefore the need to explore the application of the speech recognition task is seen in the following section of this research work.

It is possible to say that Continuous speech is a set of complicated audio signals which makes producing them artificially difficult. Speech signals are usually considered as voiced or unvoiced, but in some cases they are something between these two. Voiced sounds consist of fundamental frequency and its harmonic components produced by vocal cords. The vocal tract modifies this excitation signal causing formant and sometimes anti-formant frequencies (Witten 1982).

1.1.1. Background

Speech and hearing have involved as a main tool of communication among human beings (Luh, 2004). The basic building block of the speech of any language is a set of sounds called phonemes. Starting from our childhood we learn the skill of communication naturally till we grow up which is very complicated. Even with the difference in terms of accent, articulation, nasality, roughness, volume, pitch,

pronunciation, and speed, we are still able to interpret the speech most of the time as long as the spoken language is the language that we are familiar with (Luh, 2004).

Due to the familiarity to spoken language, we would also hope to interact with machines via speech. Scientists and researchers are finding their ways to produce an efficient speech recognizer so that a natural human –machine interface could be invented that replace the primitive interfaces, such as keyboard and mouse for the computer (Luh, 2004). With the existence of this human-machine interface, valuable applications would come into our life to make jobs done easier and effectively.

Because of the attractiveness of designing an intelligent machine that can recognize the spoken language, studies have been done in various fields to achieve this goal. From the process of speech production and perception in human beings to the way human brain learns to speak and listen, domain expertise and knowledge from wide range of disciplines are required for successful speech recognition systems.

1.1.2. Types of speech recognition

According to (Cook, 2002), speech recognition systems can be separated into several different classes by describing what types of utterances have the ability to recognize. These include isolated word, connected words, continuous or spontaneous speech recognition. This classification is based on the fact that one of the difficulties of Automatic speech recognizer (ASR) is the ability to determine when a speaker starts

and finishes an utterance (Cook, 2002). Most packages can fit into more than one class, depending on which mode they are using.

On the other hand some authors classify speech based on the characteristics of speech during their utterances. Accordingly, Kurzweil, (2002) classified in the following manner:

a. Continuous Versus Discrete Speech

Speech recognition systems are generally classified as discrete or continuous systems that are speaker dependent or independent. **Discrete systems** maintain a separate acoustic model for each word, combination of words, or phrases and are referred to as Isolated word Speech Recognition (ISR). Continuous speech recognition systems, on the other hand, respond to a user who pronounces words, phrases, or sentence that are in a series or specific order and are dependent on each other, as if linked together (Kurzweil, 2002).

Some speech systems only need identify single words at a time (e.g., speaking a number to route a phone call to a company to the appropriate person), while others must recognize sequences of words at a time. The isolated word systems are, not surprisingly, easier to construct and can be quite robust as they have a complete set of patterns for the possible inputs. Continuous word systems cannot have complete representations of all possible inputs, but must assemble patterns of smaller speech events (e.g., words) into larger sequences (e.g., sentences).

b. Speaker Dependent versus Speaker Independent

A speaker dependent system requires that the user record an example of the word, sentence, or phrase prior to its being recognized by the system; that is, the user trains the system. Some speaker-dependent systems require only that the user record a subset of system vocabulary to make the entire vocabulary recognizable. A speaker-independent system does not require any recording prior to system use. Instead, when a user identifies him or she, a speaker adaptive system adapts the word, sentence, or phrase to the user's voice as the user corrects recognition errors (ibid).

c. Small versus Large Vocabulary Systems

The other classification according to (Kurzweil, 2002) is based on the vocabulary size involving small versus large and the third categories called medium vocabulary size. The Small vocabulary systems are typically less than 10 words and it is possible to get quite accurate recognition for a wide range of users. While a large vocabulary system on the other hand constitutes more than a vocabulary size of 1000 words and or phrases typically need to be speaker dependent to get good accuracy. Finally, there are also mid-size systems or medium size vocabulary, which lies in between the aforementioned vocabulary size.

Accordingly, based on the above discussion, this research is mainly emphasis on the continuous speech or continuous utterance, speaker independent and medium vocabulary. In addition to this in the following section we will discuss the core issues about automatic speech recognition technologies so as to have deep understanding about the subject matter under investigation.

1.2. Automatic Speech Recognition (ASR) - Evolution

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 (the basic sounds of the language) and tries to explain how they are acoustically realized in spoken utterance (Juang and Robiner, 2004).

All these elements include the phonemes and the corresponding place and manner of articulation used to produce the sound in various phonetic contexts. Bell Laboratories also built in the early 1950's the system for isolated digits recognition for a single speaker (Juang and Robiner, 2004), using the formant frequencies measured (or estimated) during vowel regions each digits.

Furthermore the works of Sakai and Doshita(1962) in the University of Kyoto as cited by *Rabiner(2004)* involved the first use of a speech segmenter for analysis and recognition of speech in different portions of the input utterances. In contrast, an isolated digit recognizer implicitly assumes that the unknown utterance contained a complete digit (and no other speech sounds or words) and thus did not need an explicit "segmenter". Finally the Kyoto University's work could be considered a precursor to a continuous speech recognition system.

Presently the art of state speech recognition is grown to the various domain of recognizing audio speeches, video speeches, telephone and the like. But as it is also language dependent already indicated in different literatures, we need to navigate through different literatures. Therefore for Afaan Oromoo is also one of those languages to be examined highly with respect to speech recognition.

On the other hand viewing the overall speech recognizer how it works is usually the road to our destination. Basically the automatic speech recognizers have the following components as shown in the diagram from figure 1.1.

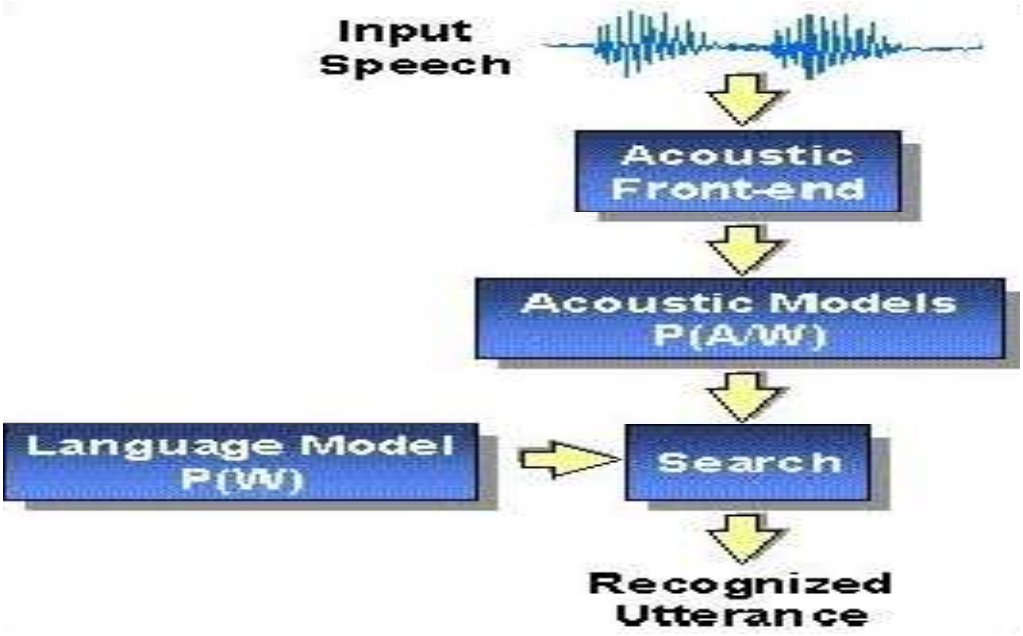

Figure 1.1. An overview of speech recognition system (Hualin Gao, et. al , 2002).

In this hypothetical example, the speech has two main parts namely the acoustic model (P(A/W))and the language model (P(W)). To build both models for the given language, there are pre-processing performed starting from recording and transcribing and preparing control file that will help for generating the acoustic and language models. Up on this the speech recognizer model need to search for the utterance that best fit to the test data submitted for the recognizer. Furthermore the fifth chapter will elaborate these issues in detail.

1.3. Statement of the Problem and Justification

The needs for designing a tool that facilitates the interaction of machine with human are underway from its early attempts to the current state-of-the-art. Ashenafi, (2009), indicated natural speech is considered to be easier than other mechanisms such as typing, pressing, rolling and sliding. One of the reasons for such ease of use is that it is trivial to observe human beings learning the speaking skill before any of the aforementioned skills used as man machine interaction mechanism.

Rabiner and Juang (1993) stated that speech recognition is rich field of research for both practical and intellectual reasons. With this regard, the use of speech to communicate between human beings is inevitable of the primary choice and with no substitutes. This is because there are physically disabled peoples, which means handicapped peoples, who use the language with one or the other way round lacked the means to interact with the machine. On this regard some people have difficulty typing due to physical limitations such as Repetitive Strain Injuries (RSI)[1], muscular dystrophy, and many others. For instance, people with difficulty hearing could use a system connected to their telephone to convert the caller's speech to text. In addition to this the domain is reach in variety of application such as dictation, command and control, telephony.

Hence some local researchers tried to alleviate the problem of speech in general and speech recognition specifically that the physical disabled person faces. To mention,

[1] occupational overuse syndrome affecting muscles, tendons and nerves in the arms and upper back

Solomon (2001), Kinfe(2002), Zegaye(2003), Hussien, (2004), have undertaken their studies on isolated word speech recognizer and related issues for specifically Amharic language using Hidden Markov Model(HMM) and Hafte(2009) for Tigrigna.

But apart from the Amharic that is the working language of the Federal Democratic Republic of Ethiopia; Afaan Oromoo is one of the languages predominantly spoken in Ethiopia and some neighbouring countries like Kenya and some part of Somalia constituting more than 40 million people (Asafa, 2010). In addition to these Afaan Oromoo is one of the languages used for instructional purpose from elementary to university programs and working language of the Regional State of Oromiya, and used by many people in the country due to the large geographic coverage and number of speakers.

To the best knowledge of the researcher, there is only one research undertaken by Ashenafi, (2009) toward designing isolated word speech recognition for Afaan Oromoo. He attempts to develop isolated word speech recognizer by taking 50 Afaan Oromoo words each with 20 utterances and arrive nearly at 82% accuracy using HMM technique and the sphinx4 tools. However, recognizing human speech, specifically continuous (connected) speech is necessary to pave the way for the full working recognizer in the language.

Furthermore, Ashenafi, (2009) also recommend the need for further research to design a continuous speech recognizer for Afaan Oromoo so as to increase the performance of the ASR system. Accordingly, this research is initiated to investigate

the possibility of developing a continuous speech recognizer for Afaan Oromoo language.

1.4. Objective of the Study

1.4.1. General objective

The general objective of this research is to explore the possibility of developing a Continuous, speaker independent Automatic Speech Recognizer for Afaan Oromoo that enables Afaan Oromoo speech utterances to be transcribed in the computer understandable text.

1.4.2. Specific objectives

In order to achieve the general objectives, a set of specific objectives are set. These specific objectives include:

- Perform a comprehensive literature review of both related to automatic speech recognizer and the language under study
- Explore the available speech recognition tools and techniques that are appropriate for this research
- Collect sample phrase and /or short statements, digitize and perform feature extraction with the appropriate tools.
- Select algorithm and models suitable for developing language models, acoustic model and pronunciations lexicon.
- Design a prototype Afaan Oromoo continuous speech recognition system

- Evaluate the performance of the recognizer using appropriate accuracy tracker.
- Forward conclusion and recommendations for further research in the area.

1.5. Scope and limitation of the study

The research aims at checking the applicability and possibility of developing a prototype for Afaan Oromoo continuous speech recognizer. The system also integrated natural language processing and signal processing the speech signal in to computer understandable text. Language model, acoustic model and pronunciation dictionary are the main part of the recognizer.

Due to limitation of time and unavailability of Afaan Oromoo corpus for the research, the researcher is limited in working with 70 selected Afaan Oromoo phrases and simple sentences.

In this thesis work the dialect differences were not considered within the language Afaan Oromoo. This is mainly because the variation in dialect causes the differences in the performance of the recognizer and need to perform a lot of acoustic tasks..

Furthermore, the researcher also encounters some difficulties during recording the speech data from selected individual due to unwillingness of the respondents and lack of uttering correctly the given phrases and sentences, as one individual are expected to utter all the phrases. In addition to this the recording environment needs quit environment and in most cases it is difficult to get respondents in the specified areas.

1.6. Methodology

In order to achieve the objective of this research, the following methods were employed during the course of implementing a continuous, speaker independent speech recognizer for Afaan Oromoo.

1.6.1. Literature Review

Comprehensive literature review were performed form books, journals, and the internet in order to investigate the underlying principles/ theories of the various approaches, techniques and tools that were employed in the research. Further literature review is also performed on the language Afaan Oromoo and the phonetic characteristics of the language as well. Besides, literatures referring to similar research in the area of speech recognition were reviewed. This helped the researcher to understand the problem and to support the result in a scientific manner to be empirically evaluated.

1.6.2. Data Selection and Collection

By consulting the domain expert on both Afaan Oromoo language and phonetics the researcher studied the characteristics of the language under study. Accordingly the data constituting all phonetic characteristics of Afaan Oromoo language were selected and collections of those data were performed so as to reach the appropriate result in the language recognizer.

Since the sound of sub-word unit is also influenced by its context, the speech data should include all phonemes as many as different phonetic contexts as possible. In practice this means that, depending on the quality of the data and the complexity of the acoustic models, about ten to thirty thousand phonetically rich sentences are necessary (Wiggers, 2001). Therefore the researcher used the selected words, phrases and short sentences that constitute all phonetic characteristics of the language Afaan Oromoo. To achieve this, 70 words/phrases/short sentences are selected in terms of their utterances, by consulting the domain expert which is uttered by 30 selected persons and students constituting a total of 2100 utterances.

1.6.3. Development Tools and Techniques

In the development process of the prototype the researcher used appropriate tools and techniques indicated in different literatures. Specifically in this research the researcher used the machine learning techniques appropriate for speech in comparison from the available machine learning approaches such as ANN, HMM and SVM the recognizer. Tools like Sphinxtrain for training and Sphinx4 for modelling the system are dominantly used in different literatures and accordingly discussed latter. Therefore, for this specific research, HMM and the tool from sphinx system (Sphinxtrain and Sphinix4) was used.

Sphinx System is an open source software used by developers and some academic and research institutes. With the platform independent, well structured decoder Sphinx4 for real applications, its importance has greatly increased (Juraj Kačur , 2004). In addition to

this in the literature and methodology parts it is well justified why this techniques and tools are selected for this research.

1.6.4. Evaluation and Testing Procedures

As clearly indicated in different literatures, the testing activities are one of the important tasks to be performed while doing one experiment. Therefore, in this research the performance is evaluated against the test speech data set prepared using the sphinx4 decoding tool. Here the classification and other preparation of the training and test data sets were performed manually. Whereas the training received from the sphinx train was then taken to the decoder and the task of measuring the performance is done automatically using the appropriate scripts for the selected words, phrases and sentence/statement to check the applicability of the recognizer by evaluating its performance. As Zegeye, (2003) also indicated the most important and common testing parameter used to evaluate speech recognition system is accuracy of recognition. The performed evaluation is conducted for checking the level of uncertainty in the given grammar of the corpus under investigation. Furthermore, the performance of the evaluation is depicted in chapter five of this research paper latter for its applicability and usability in the language.

1.7. Significance of the study

Conducting research on automatic speech recognition system for Afaan Oromoo is useful in many ways. One of these is investigating the possibility of developing continuous speech recognizer and which will pave the way in identifying the possible way of developing an automatic speech recognizer for Afaan Oromoo. This in turn could be extended to improve the man computer interaction particularly pertaining to the speaker of Afaan Oromoo Language.

As Cook (2003) has clearly stated the general application of an automatic speech recognizer as dictation, command and control, telephony and use for disable people[2]. These applications of automatic speech recognizer could be extended for Afaan Oromoo speech recognizer as well. In addition to the above benefit that the research gives, it also uses as tool for the researchers to learn more how the area is performing. Furthermore this study might pave ways for further research and indicate the applicability of the models and tools for the future real development of the recognizer for the language.

Beyond these, the research work will have high benefit when it is looked from different angles. For example, the attempt towards developing the recognizer will be enhanced and will pave ways for the upcoming young researcher. In addition to this the researcher also gets adequate knowledge towards the discipline.

[2] Disabled people refer to people with physical disability including handicapped.

1.8. Organization of the Thesis

The thesis is organized in six chapters of which are briefed in the following ways:

The first chapter or chapter one is mainly discussing about the introduction, statement of the problems and its justification, the objectives of this research, the methodologies to follow in brief, the beneficiaries, the scope and limitation of the research.

The second chapter all in all discussed the underlying principles and theories of speech technologies, from speech production to the characteristics of speech production places and the different machine learning approaches including HMM, ANN and SVM and the techniques and methods used in speech technologies in general.

The third chapter mainly discusses literature of the people and language under investigation. The researcher in this part tried to illustrate the language characteristics and the speech nature and the vowel and consonants of the language apart from the nature of noun and the verbs. This is to include all the phones of the language under investigation for continuous speech of Afaan Oromoo.

The fourth chapter is all about the methodologies used for this specific research in generally and about HMM in particular. It also indicates in detail about the tools and techniques employed in this research.

The fifth chapter is all about the design and experimentation of the recognizer under consideration. Therefore in this chapter the researcher tried to show the analysis of the experimentation in line with the discussion of the results obtained from the tool.

The six chapter presents the conclusions arrived in and the recommendations made by the researcher so as to fully deploy the recognizer and for further researches to be conducted latter in this domain.

CHAPTER TWO

SPEECH AND LANGUAGE

Introduction

In this chapter we covered things related speech technologies from its production to recognition. The characteristics of those places of production of sounds along with tools and techniques used reviewed for this thesis work. Furthermore some local and international research works related to speech recognition were discussed and upon that the appropriate tools and techniques are selected.

2.1 Human Speech Production Systems

A speech production model that more directly simulates the physical process of human speech production comprises lungs, vocal cords, and the vocal tract (Campbell, 1997). The vocal cords are expressed as a simple vibration model, and the pitch[3] of the speech changes according to adjustments in the tension of the vocal cords. When the vocal cords close, their vibration results in voiced sounds; when they open, this vibration stops, and unvoiced sounds result.

Therefore it is important to note that speech is achieved by compression of the lung volume causing air flow which may be made audible if set into vibration by the activity of the larynx. This sound can then be made into speech by various modifications of the supra laryngeal vocal tract as indicated in the figure 2.1. In

[3] represents the perceived fundamental frequency of a sound

addition to this the production of speech in most cases performed in the following processes.

a. Lungs provide the energy source - **Respiration**
b. Vocal folds convert the energy into audible sound - **Phonation**
c. Articulators transform the sound into intelligible speech - **Articulation**

Therefore it is possible to elaborate the production and, conversion and transformation of speech in human from those processes view point. Accordingly let us examine what those three terminologies so as to improve our understanding

2.1.1. Respiration

In normal speech, the action of the respiratory apparatus during exhalation[4], this provides a continuous stream of air with sufficient volume and pressure to initiate phonation. The stream of air is modified in its course from the lungs by the facial and oral structures, giving rise to the sound symbols that are recognized as speech.

2.1.2. Phonation

Phonation has slightly different meanings depending on the subfield of phonetics. Among some phoneticians, *phonation* is the process by which the vocal folds produce certain sounds through quasi-periodic vibration. This is the definition used among those who study laryngeal anatomy and physiology and speech production in

[4] movement of air out of the bronchial tubes, through the airways, to the external environment during breathing.

general. Other phoneticians, though, call this process quasi-periodic vibration voicing, and they use the term phonation to refer to any oscillatory state of any part of the larynx that modifies the airstream, of which voicing is just one example. As such, voiceless and supra-glottal phonation is included under this definition, which is common in the field of linguistic phonetics.

2.1.3. Articulation

When sound is produced at the larynx, that sound can be modified by altering the shape of the vocal tract above the larynx. The shape can be changed by opening or closing the velum (which opens or closes the nasal cavity connection into the oropharynx), by moving the tongue or by moving the lips or the jaw. After all the main organs of the human speech production system articulators are illustrated in the following diagram showing the participating organs in their respective names.

Figure2.1. An overview diagram of the major articulators. (Extracted :Mannell, 2001)

2.2. Basic Terms and Concepts of Speech Recognition

Following are a few of the basic terms and concepts that are fundamental to speech recognition. It is important to have a good understanding of these concepts when developing any speech recognizer. These words are extracted from the speech recognition white paper Speech production HOWTO (Cook, 2003)

Utterances

When the user says something, this is known as an utterance**.** An utterance is any stream of speech between two periods of silence. Utterances are sent to the speech engine to be processed.

On the other hand silence, in speech recognition, is almost as important as what is spoken, because silence delineates the start and end of an utterance.

Utterances are sent to the speech engine to be processed. If the user doesn't say anything, the engine returns what is known as a silence timeout - an indication that there was no speech detected within the expected timeframe, and the application takes an appropriate action, such as re-prompting the user for input. An utterance can be a single word, or it can contain multiple words (a phrase or a sentence).

Pronunciations

The speech recognition engine uses all sorts of data, statistical models, and algorithms to convert spoken input into text. One piece of information that the speech recognition engine uses to process a word is its pronunciation, which

represents what the speech engine thinks a word should sound like. Therefore we need to prepare a pronunciation dictionary that the speech recognizer uses during recognizing the continuous Afaan Oromoo speech.

Grammars

Specifying the words and phrases that users can say to your application. These words and phrases are defined to the speech recognition engine and are used in the recognition process. A grammar uses a particular syntax, or set of rules, to define the words and phrases that can be recognized by the engine. A grammar can be as simple as a list of words or phrases, or it can be flexible enough to allow such variability in what can be said that it approaches natural language capability.

Accuracy

It is typically a quantitative measurement and can be calculated in several ways. Arguably the most important measurement of accuracy is whether the desired end result occurred or not. Another measurement of recognition accuracy is whether the engine recognized the utterance exactly as spoken. This measure of recognition accuracy is expressed as a percentage and represents the number of utterances recognized correctly out of the total number of utterances spoken. It is a useful measurement when validating grammar design.

Recognition accuracy is an important measure for all speech recognition applications. It is tied to grammar design and to the acoustic environment of the user. Therefore, we need to measure the recognition accuracy for our recognizer, and may want to

adjust our recognizer and its grammars based on the results obtained when you test your application with typical users. Therefore, this thesis work will be conducted on the bases of the above facts.

2.3. How Speech Recognition Works

Speech recognition engine has the role of converting raw audio input it to recognized texts. As shown in the diagram below, the major components (high level) of the speech recognition models are:

- ✓ Audio input (speech input)
- ✓ Grammar(s) (Lexical model)
- ✓ Acoustic Model
- ✓ Language model
- ✓ Recognized text

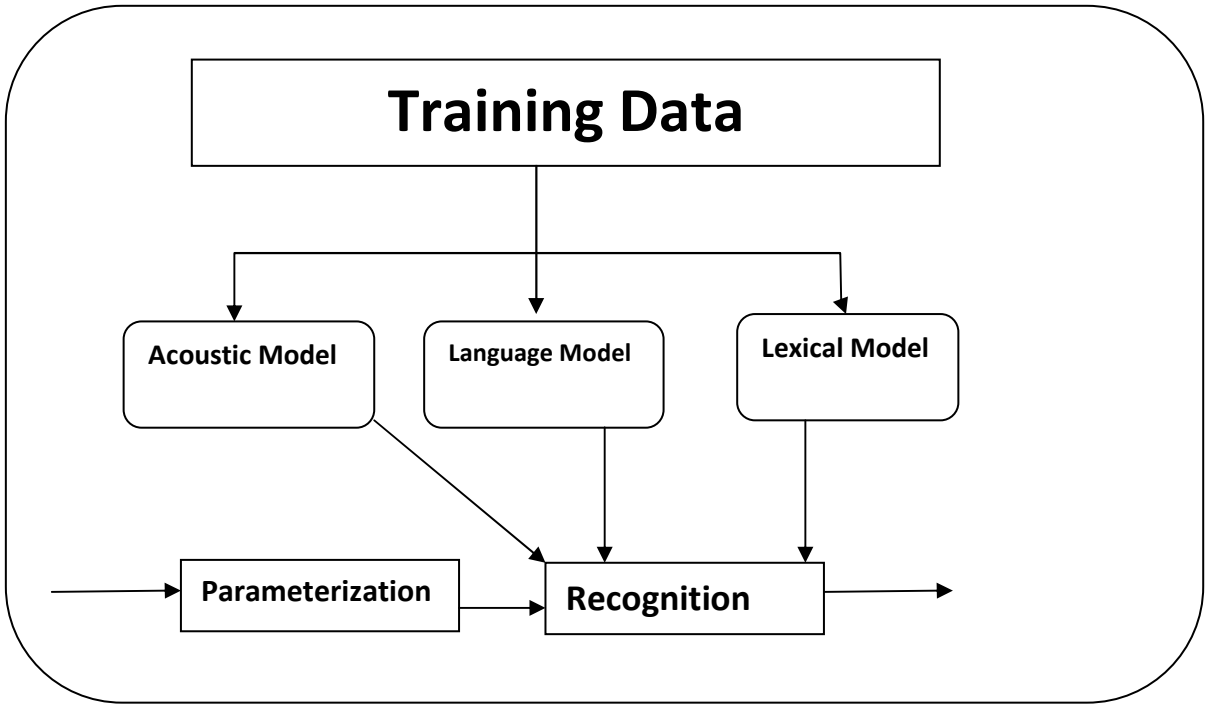

Figure 2.2: Major Components of Speech Recognition (from: Zue et al., 1996)

The first thing that we want to take a look at is the audio input coming into the recognition engine. Of course in the course of recoding the audio files the Afaan Oromoo text corpus was prepared so as to be uttered by reading by selected individuals. It is important to understand that this audio stream is rarely pristine (Kurzweil, 2002). It contains not only the speech data (what was said) but also background noise. This noise can interfere with the recognition process, and the speech engine must handle (and possibly even adapt to) the environment within which the audio is spoken.

It is also important to note that the job of the speech recognition engine is to convert spoken input into text. To do this, it employs all sorts of speech data, statistics, and software algorithms. Its first job is to process the incoming audio signal and convert it into a format best suited for further analysis (feature files). Once the speech data is in the proper format, then the engine searches for the best match. It does this by taking into consideration the words and phrases it knows about (the active grammars), along with its knowledge of the environment in which it is operating. The knowledge of the environment is provided in the form of an acoustic model. Once it identifies the most likely match for what was said, it returns what it recognized as a text string.

Most speech engines try very hard to find a match, and are usually very tolerant. But it is important to note that the engine is always returning its best guess for what was said.

2.4. History of Automatic Speech Recognition System

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 (the basic sounds of the language) and tries to explain how they are acoustically realized in a spoken utterance(*Rabiner,2004*).

These elements include the phonemes and the corresponding place and manner of articulation used to produce the sound in various phonetic contexts. For example, in order to produce a steady vowel sound, the vocal cords need to vibrate (to excite the vocal tract), and the air that propagates through the vocal tract results in sound with natural modes of resonance similar to what occurs in an acoustic tube (*Rabiner, 2004*). These natural modes of resonance, called the *formants* or *formant frequencies,* are manifested as major regions of energy concentration in the speech power spectrum. According to (K. H. Davis, 1952) in 1952, Davis, as sited by(*Rabiner, 2004*) Biddulph, and Balashek of Bell Laboratories built a system for isolated digit recognition for a single speaker, using the formant frequencies measured (or estimated) during vowel regions of each digit.

Later on in the 1960's, (Forgie, 1956) as sited by(*Rabiner, 2004*) Forgie built a speaker-independent 10-vowel recognizer. Several Japanese laboratories also demonstrated their capability of building special purpose hardware to perform a speech recognition task. Among all the most notable were the vowel recognizer of Suzuki and Nakata at the Radio Research Lab in Tokyo, the phoneme recognizer of Sakai and Doshita at Kyoto University (Sakai and Doshita, 1962) as cited by *Rabiner, 2004*.

Furthermore the work of Sakai and Doshita involved the first use of a speech segmenter for analysis and recognition of speech in different portions of the input utterance. 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.

In addition to the above attempt, the IBM and AT&T Bell Laboratories approaches to speech recognition both had also profound influence in the evolution of human-machine speech communication technology of the last two decades. One common theme between these efforts, despite the differences, was that mathematical formalism and rigor started to emerge as distinct and important aspects of speech recognition research. While the difference in goals led to different realizations of the technology in various applications, the rapid development of statistical methods in the 1980's, most notably the HMM framework (Rabiner, 1983) caused a certain degree of convergence in the system design. Today, most practical speech recognition systems are based on the statistical framework and results developed in the 1980's, with significant additional improvements in the 1990's.

The developments of automatic speech recognizer is this particular thesis is performed using the HMM which is prominent with statistical approach and thus let us examine the statistical approaches in detail and other approaches in speech recognition in the next session so as enhance the approach to be used.

2.4.1 Speech Recognition Approaches

The different automatic speech recognition (ASR) systems are based on different kind of approaches. ASR approaches can be divided into different categories, but usually hybrid methods are applied. In general, there are three classical approaches:

1. The acoustic-phonetic approach
2. The statistical pattern recognition approach and
3. The artificial intelligence (AI) approach.

The acoustic-phonetic method is the oldest speech recognition approach originating from the 1950's, the AI approach is the youngest and least known. Statistical methods are by far most commonly applied in modern speech recognizers. Let us examine one by one so as to get insight of the three approaches in the following section.

2.4.1.1 Acoustic-Phonetic Approach

The acoustic-phonetic (AP) approach is based on the theory of acoustic phonetics that postulate that there exist finite, distinctive phonetic units in spoken language and that the phonetic units are broadly characterized by a set of properties that are manifest in the speech signal, or its spectrum, over time. The approach assumes that the rules governing the phoneme variability are relatively simple and easily learnable.

According to (Rabiner & Juang 93) the AP approach to speech recognition performs the following tasks in different phases:

The first step is the speech analysis system, which provides an appropriate (spectral) representation of the characteristics of the time-varying speech signal. The most common techniques of spectral analysis are Discrete Fourier Transform (DFT), Linear Predictive Coding (LPC), or Mel-Scaled Frequency Cepstral Coeffients (MFCC).

The next step is the feature-detection stage. The spectral measurements are converted in a parallel fashion to a set of features describing the broad acoustic properties of the various phonetic units, e.g. nasality, frication, format locations, voiced/unvoiced classification, and energy ratios.

The third step is the heart of the AP recognizer: the segmentation and labelling phase, in which the system tries to find feature stable regions and then label those regions according to how well the features within that region match those of individual phonetic units. The result of the this step is usually a phoneme lattice from which a lexical access procedure determines the best matching of word or sequences of words, as shown in Figure 2.3.

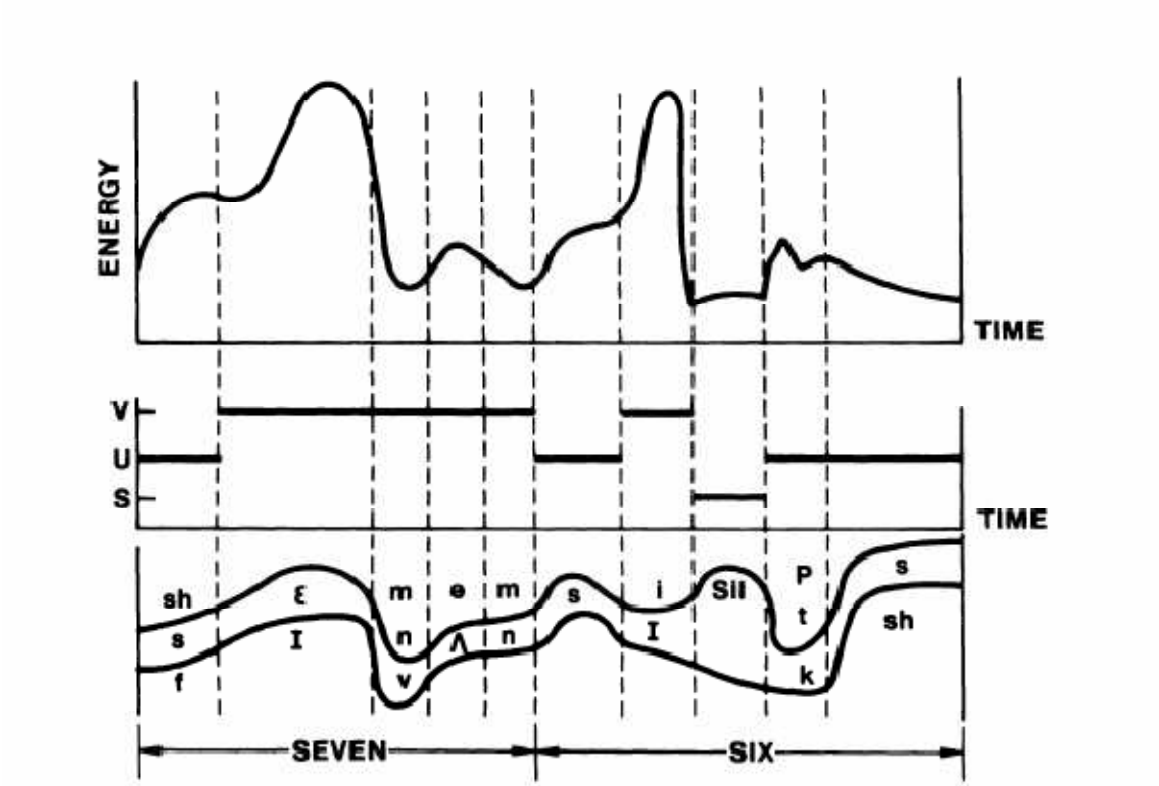

Figure 2.3: Segmentation and labelling for word sequence "seven-six"(Rabiner & Juang, 93)

The AP approach has several dilemmas that hinder its functionality as an ASR system. The method requires extensive knowledge of the acoustic properties of phonetic units. This knowledge is, at best incomplete, and at worst totally unavailable for but the simplest of situations (e.g. steady vowel). Also, the features are often based on non-optimal and intuition based ad hoc considerations and the design of the sound classifiers is also non-optimal. Furthermore, there is no well-defined, automatic procedure for tuning the method (i.e. adjusting the decision threshold) on real, labelled speech. Moreover, there is no standard linguistic way of labelling the training speech. Naturally, these problems need to be solved before the approach can be well utilized in practice.

Due to the above limitations, the AP approach has not achieved the same success in practical systems as have alternative method. But its underlying ideas are still used in the artificial intelligence based recognizers.

2.4.1.2 Statistical Pattern Recognition Approach

The statistical approaches are by far most commonly applied in modern speech recognizers due to its strength discussed here.

The pattern recognition paradigm has four steps, namely (Rabiner & Juang 93): Feature extraction**,** in which the important features is extracted from the input signal and represent it a form of feature pattern. The feature extraction techniques include DFT, LPC, LPCC, MFCC and etc.

(1) **Pattern training,** in which one or more test patterns corresponding to speech sounds of the same class are used to create a pattern representative of the features of that class. The resulting pattern, generally called reference pattern, can be template, derived from some type of averaging technique, or it can be a model that characterizes the statistics of the features of the reference pattern.

(2) **Pattern classification,** in which the unknown test pattern is compared with each class reference patterns, and a measure of similarity between the test pattern and each reference pattern is computed.

(3) **Decision logic,** in which the reference pattern similarity scores are used to decide, which reference pattern (or possibly which sequence of reference patterns) best matches the unknown test pattern.

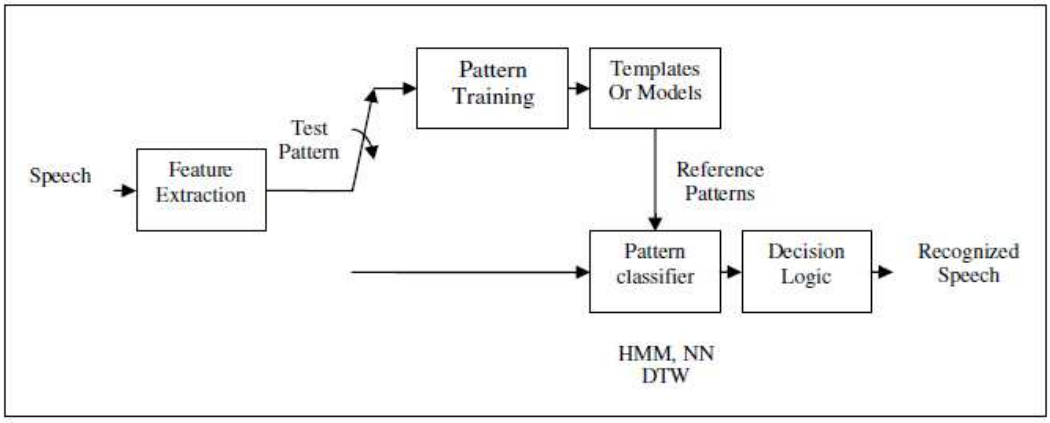

Figure 2.4: Block diagram of pattern recognition speech recognizer (Rabiner & Juang, 93).

The general **strengths** and **weaknesses** of the statistical pattern recognition include the following:

(1) The performance of the system is sensitive to the amount of training data available for creating sound class reference patterns; generally the more training, the higher the performance of the system for virtually any task.

(2) The reference patterns are sensitive to the speaking environment and transmission characteristics of the medium used to create the speech; this is because the speech spectral characteristics are affected by transmission and background noise.

(3) No speech-specific knowledge is used explicitly in the system; hence, the method is relatively insensitive to the choice of vocabulary words, task syntax, and task semantics.

(4) The computational load for both pattern training and pattern classification is generally linearly proportional to the number of patterns being trained or recognized; hence, computation of a large number of sound classes could and often does become prohibitive.

(5) Because the system is insensitive to sound class, the basic techniques are applicable to a wide range of speech sounds, including phrases, whole words, and sub-word units. A basic set of techniques developed for one sound class (e.g. words) can generally be directly applied to different sound classes (e.g., sub-word units) with little or no modifications to the algorithms.

(6) It is relatively straightforward to incorporate syntactic (and even semantic) constraints directly into the pattern recognition structure, thereby improving recognition accuracy and reducing computation.

The factors that distinguish different pattern recognition approaches are the types of features, the choice of templates or models for reference patterns, and the method use to create reference patterns and classify unknown test patterns. In template based approach, let define a test pattern, T, as the concatenation of spectral frames over the duration of the speech and in a similar manner a set of reference patterns $\{R_1, R_2, ..., R_V$ where each reference pattern R_j, is also a sequence of spectral frames. The pattern comparison stage of this approach is to determine the dissimilarity or distance of T to each of the R_j , $1 \pounds j \pounds V$, in order to identify the reference pattern that has the minimum dissimilarity, and to associate the spoken input with the pattern. The Dynamic Time Warping (DTW) technique is a kind of template matching

approach used in speech recognition (Sakoe & Chiba 1978) as referred by Ming (2007). The pattern comparison concept is extended to the case of using statistical model. The test pattern is compared to a mathematical or statistical characterization of each reference pattern, rather than to a specific reference pattern sequences. HMM is a well known and widely used statistical method of characterizing the spectral properties of the frames of speech pattern (Rabiner, 1989).

The template based approach has two major drawbacks compared to statistical based approach. First is their incapability of model acoustic variability, except in a coarse way by assigning multiple of reference tokens are used to characterize the variation among different utterances.

Secondly, in practice the template approach is limited to word-unit, because it is hard to record or segment a sample shorter than a word- so templates are useful only in small systems which can afford the luxury of using word-models. The statistical model can otherwise represent both word and sub-word units which allow its advantages in large vocabulary system. Due to this and some other parameters to be discussed latter the statistical approach of HMM is used in this thesis for Afaan Oromoo continuous speech recognition.

2.4.1.3 Artificial Intelligence Approach

The artificial intelligence (AI) approach is a hybrid of the acoustic-phonetic and statistical recognition methods. It tries to mimic the human intelligence in visualizing, analyzing and decision making progress on the measured acoustic features. The main idea of AI is to collect and employ knowledge from a number of

sources for solving the problem in question. The knowledge sources (KS) are wide-ranging from fields of acoustic, lexical, syntactic, semantic and pragmatic knowledge (Rabiner & Juang 1993).

Most important techniques in this approach is the use of an *expert system* for segmentation and labelling of the acoustic signal, learning, and adaptation over time, the use of *artificial neural network* (*ANNs*) for distinction between similar sound classes and learning the relations between all known inputs and phonetic data. The neural network could represent a separate structural approach to speech recognition or regarded as an implementation architecture possibly incorporated in any of the three classical speech recognition approaches. The oldest examples are applications combining ANNs with conventional technologies originate from the late 1980s applying, e.g. vowel, word and digit recognition using *multi-layer perceptrons* (*MLPs*), *learning vector quantization (LVQ*) and *time delay neural network* (*TDNNs*). The modern methods are hybrids of ANNs and HMMs applying, e.g. recurrent neural networks (RNNs), self-organizing maps (SOMs) and mixtures of experts (Rabiner & Juang 1993).

2.5. Modelling and Classification Techniques in Speech Recognition

The main modelling and classification techniques currently used for ASR are described in this section.

2.5.1 Dynamic Time Warping

Dynamic time warping (DTW) is one of the oldest and most important algorithms in speech recognition (Sakoe & Chiba 1978) as referred by Ming (2007). 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. The algorithm finds the optimal nonlinear alignment between the unknown speech patterns with the reference pattern which both may vary in duration due to different speaking rate to obtain their global distance which indicate their similarity.

While effective in pattern recognition, what the DTW algorithm lacks is the long processing time and large pattern 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. Besides, DTW approach is limited to word template.

Ney (1984) has extended the usage of DTW in isolated word to continuous speech recognition with the algorithm called One Stage DTW. Here the goal is to find the

optimal alignment between the speech sample and the best sequences of reference words.

Although DTW is computationally much simpler than HMM, its computational 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 and continuous speech recognition (Wong, 1998).

Therefore the DTW techniques have been generally over masked by the more powerful and flexible HMM models and for this reasons the researcher selected the HMM technique.

2.5.2 Hidden Markov Model

In most current speech recognition systems, the acoustic modelling components of the recognizer are almost exclusively based on HMMs. The ability to statistically model the variability in speech has been the main reason for the success that HMMs have enjoyed over years. HMMs provide an elegant statistical framework for modelling speech patterns using a Markov process that can be represented as a state machine. The temporal evolution of speech is modelled 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.

Basically, a HMM model 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.

The HMM technique is selected for this specific research and the studied in detail in separate chapter for further investigation and its applications in line with the basic assumptions and probability functions. Therefore so as to justify well we need to see the other popular machine learning techniques so as to make the selection justifiable.

2.5.3 Neural Network

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

Besides, they are easy to implement. 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 modelling of temporal structure whereas the NNs are typically formulated as classifier of static data.

In addition to this, the estimation process of NNs is significantly more computationally expensive than HMMs. The relative strength of NN and HMM approaches have stimulated much research. TDNN and RNN have been investigated

as a way to improve the NN's use of contextual information. NNs have been used in parallel with HMMs as a second classifier in speech recognizer (Devillers & Dugast 1993). Perhaps the most important development has been the use of hybrid HMM/NN architectures. Hybrid HMM/NN approaches aim to combine the best features of both architectures. In this case the NNs are embedded in a HMM framework (Tebelskis 1995).

The performance of these hybrid systems have been competitive with many HMM-based systems and typically require a significantly reduced parameter count. The hybrid connectionist systems also provide a way to mitigate some of the assumptions made in HMM systems that we know are incorrect for the human speech process (Russell & Moore 1985). One such significant assumption is that of independence of observations across frames (Ostendorf *et al*. 1996). Hybrid systems mitigate this problem by allowing the NN classifiers to classify based on several frame of acoustic data at a time (Tebelskis 1995).

Due to the limitation of NN compared to HMM approach, the HMM approach is used in this work.

2.5.4 Support Vector Machine

The HMM and NN based speech recognition system do not yield good generalization. These can be seen from the fact that the performance of these systems on speaker dependent tasks is significantly better than on speaker-independent tasks.

The SVM is one of machine learning technique which provides better generalization and convergence stimulates research on applying it on speech recognition.

SVM are founded on the principle of Structural Risk Minimization (SRM) and the result of SRM is a classifier with the least expected risk on the test set and hence good generalization. SVMs in their simplest form are hyper plane classifiers. The power of SVMs lies in their ability to implicitly transform data to a high dimensional space and to construct a linear binary classifier in this high dimensional space. Since this is done implicitly, without having to perform any computations in high dimensional space, neither the dimensionality of the data nor the sparsity of the data in the high dimensional space is a problems with SVMs. The hyper-planes in the high dimensional transform space result in complex decision surfaces in the input data space (Ganapathiraju, 2002).

SVMs have been applied successfully on several kinds of classification problems and have consistently performed better than other non-linear classifier like neural network and mixtures of Gaussians. The development of efficient optimization schemes led to the use of SVMs for classification of larger tasks like text-categorization.

SVMs are not designed to handle temporal structure of data and thus not suitable in modelling speech data which evolve with time. To take account of the temporal variation of speech data, the SVM/HMM hybrid has been proposed (Ganapathiraju 2002), in which SVM is used as are used as part of a post-processing stage. This hybrid approach uses SVMs to process information supplied by a baseline HMM

system to arrive at the final hypothesis. The baseline HMM system is used to provide segmentations in order to construct the input feature vectors for the SVMs. SVMs classify data based on distances.

In order to integrate SVMs into an HMM framework, it is need to convert these distances to posterior probabilities. Several schemes have been studied to convert SVM distances to likelihoods in order to fit the SVM classifiers into the HMM-based ASR system. However, there is limited application of SVM in speech recognition. As a result of all those pros and cons of SVM approach, the approach is not used in this thesis.

Furthermore the following subtopic discusses some of the researcher work in the area which is reviewed to strengthen all the above investigated literatures. In addition the section also shows us the main tools, techniques and methodologies that the researchers used in their respective researches.

2.6. Related works

There are many researchers conducted their researches regarding speech recognition in various parts of the world so as to support specific language under study in each case. Accordingly, there are two groups of research works selected for convenience of reviewing, the first group refers to non Ethiopian origin and secondly some local researches were considered. Therefore the researchers work along with the tools used and their respective finding is the target of the review.

Abuel (1997) in his thesis entitled as *an Arabic phoneme recognition system using Hidden Markov Models* tried for the first time the continuous speech recognition for Arabic language. In his thesis, VQ approach is used as a compression technique to reduce the computational Complexity of the algorithm. The researcher also used discrete HMM for implementing for each of the 32 Arabic phonemes. The Viterbi re-estimation method is used to estimate model parameters. Performance tests were accomplished on different observation sets using this final phoneme recognition system. Results showed that the best features to represent Arabic phonemes are the weighted cepstral coefficients plus their differenced values in one observation vector (performance 74%). While the area functions is the worst feature to represent Arabic phonemes (performance 48%). The performance test for this work was not extended for continuous speech which may contain different phonemes with variation in length.

Arisoy (2005) on his thesis entitled as *A unified language model for large vocabulary continuous speech recognition of turkey* came up with a Turkish dictation system for

newspaper content transcription application. Turkish is an agglutinative language with free words order and this characteristics of the language resulted in vocabulary explosion, large number of out-of-vocabulary (OOV) words and an increased complexity of n-gram language models in speech recognition when words are used as recognition units. In this paper, alternative language modelling units like ''stems and endings'', ''stems and morphemes'', and ''syllables'' are investigated instead of ''words''. These recognition units are compared in terms of vocabulary size, coverage, bigram perplexity and speech recognition performance. A combined model is proposed which aims to produce a balance between the OOV rate and the amount of phoneme sequence constraints on recognition units. The proposed model resulted in letter error rates of approximately 28% for a speaker independent system and 20% for a speaker dependent system. These error rates are smaller compared to the traditional word-based model for newspaper content transcription application as per the findings of the researcher.

Mohammad (2006) conducted his thesis on *a vector quantization approach to isolated word automatic speech recognition*. Doing this, the researcher came up with developing an Isolated-Word Automatic Speech Recognition System based on VQ approach. Developing this system is meant to assist customers calling a university's telephone operator to respond to their enquiries in a fast and convenient way using their natural speech. Callers are assisted using their own speech inputs to select their language preference, faculty in a university and finally select the staff name they wish to contact.

To extract features from the speech signals the MFCC algorithm was applied. Subsequently, VQ algorithm based on the principle of block coding was used for all feature vectors generated from the MFCC algorithm.

A codebook was resulted from training the VQ initial codebook and experimental results showed that the recognition rate has been improved with the increase of codebook size. Simulation results showed that the codebook size of 81 feature vectors had a recognition rate exceeded 85%.

Ming (2007) in his thesis entitled as *Malay continuous speech recognition using continuous density hidden markov model* investigated the use of CDHMM for Malay ASR. In addition to this the researcher tried fill constraints of existing Malay ASR that are: speaker-dependent, small vocabulary and isolated words, and provides a basis in developing speaker-independent Malay Large Vocabulary Continuous Speech Recognition (LVCSR). HMM based statistical modelling is used in Malay speech recognition because of its robustness and powerful technique capable of modelling of speech signals. CDHMM which model the continuous acoustic space eliminates quantization error imposed by discrete HMM and according to the findings CDHMM performs better than discrete HMM in Malay speech recognition. CDHMM with mixture densities which is capable to model inter-speaker variability performs well in multi speaker task (99% in isolated words task). A connected words ASR is developed and evaluated on Malay connected digit task and has achieved reasonably good accuracy with limited training data. The sub-word unit modelling is attempted in Malay phonetic classification and segmentation on medium vocabulary Malay

continuous speech database. Experiments are conducted to investigate different feature set and mixture components. Finally the researcher found that the basic idea of HMM implemented in other language domain can be successfully applied in the Malay language domain also.

In addition to the above researchers, there are also some local or Ethiopian researchers exerted their efforts towards speech recognition for different languages of Ethiopia. Among others, let us consider the works of some of the notable researcher in the area:

Zegaye (2003) in his thesis entitled *as Hidden Markov Model Based Large Vocabulary, Speaker Independent, and Continuous Amharic Speech Recognition* addressed recognizer developed using Hidden Markov Model; and the Hidden Markov Modelling Toolkit (HTK) was used to implement it. In order to support the acoustic models, a bigram language model was constructed. In addition, pronunciation dictionary was prepared and used as an input. Since the recognizer was meant to recognize large vocabulary and continuous speech, phonemes were opted as the basic unit of recognition. However phonemes are known to be context independent units, given that the environment in which a sound is put makes a difference in the way it is pronounced. Thus after the monophone based speech recognizer was built, it was promoted to triphone based system in which the left and right contexts were considered and modelled. Besides, the mixture components of the states of the models were incremented in view of optimizing the performance of the recognizers.

Hafte (2009) on his thesis entitled as *Hidden Markov Model Based Large Vocabulary, Speaker Independent Continuous Tigrigna Speech Recognition* conducted in HTK (Hidden Markov Model Toolkit) environment.

The researcher, in his work, used database comprised of 250 utterances that are used for training and 50 sentences for testing and evaluation. The data is pre-processed in line with the requirements of the HTK toolkit. Furthermore the researcher attempted to build speech to text conversion for Tigrigna language using the statistical approach. In order to support the acoustic models, a bigram language model is constructed. In addition, pronunciation dictionary is prepared and used as an input. In addition to the monophone based speech recognizer is built, the researcher also tried triphone based system in which the left and right contexts consider and modelled.

According to the researcher, performance tests also conducted at various stages using the training and test data. As a result of this the researcher arrived at 60.20% word level correctness, 58.97% word accuracy, and 20.06 % sentence level correctness are obtained.

Ashenafi (2009) in his thesis entitled as *A Speech Recognition System for Afaan Oromo* explored the possibility of developing an automatic speech recognizer for Afaan Oromo. In his thesis the researcher considered isolated word speaker independent speech recognizer using Hidden Markov Model and an open source speech recognition toolkit called Sphinx4.

For the purpose of achieving his objective the researcher prepared a speech corpus constructed using 50 Afaan Oromo words chosen by consulting the domain expert and uttered by 20 different people providing 1000 utterances making the entire dataset. Accordingly the data set is classified in such a way that 2/3 is used for training and the remaining 1/3 for evaluation. Pre processing and constructing of the acoustic model, the language model and the pronunciation dictionary with the help of sphinx4 and HMM was done.

The performance of the recognizer is evaluated to test the recognition accuracy, transcription speed and type of errors occurred in decoding the test set. Accuracy tracker tool is used to test the performance of the system which is an integral part of the Sphinx system. Accordingly, 82.83 % and 81.081% word level accuracy is obtained for context dependent phoneme based models and context independent word based models, respectively.

Generally the review of different literatures or related works indicates that HMM is modelling technique with statistical approach is predominantly indicated as the most dominant technique in automatic speech recognition.

Natural languages differences also results in variation as to which recognition unit, front end processing operation, language modelling technique and so on of a speech recognizer is more convenient for a particular language (Abuel, 1997). Therefore, as it depends on all those factors, the need for studying the language under investigation needs critical study so as to fill the gap. As a result the characteristics and the nature of the language under investigation is studied in chapter three.

CHAPTER THREE

PHONETICS AND PHONOLOGY OF AFAAN OROMOO

3.1. The Oromoo People and the Language Afaan Oromoo

3.1.1. The Oromoo People

The Oromo's are an ethnic group found in Ethiopia, in northern Kenya and to a lesser extent in parts of Somalia. Afaan Oromoo is spoken by nearly over 40 million members, throughout the world (Asafa, 2010). Furthermore, Asafa (2010) discussed on his paper the relative number of the Oromoo people and the various dialects found within the language Afaan Oromoo. Their native language is Afaan Oromoo, which is part of the Cushitic branch of the Afro-Asiatic language family.

3.1.2. The Afaan Oromoo Language

Afaan Oromoo or Oromiffa is an Afro-Asiatic language, and the most widely spoken of the Cushitic family. It is spoken as a first language by more than 40 million Oromoo and neighbouring peoples in Ethiopia and Kenya (Asafa, 2010).

About 95 percent of Oromo speakers live in Ethiopia, mainly in Oromiya region (Tilahun, 2002). It is also important to note that there are second language speaker which are Oromoo and non Oromoo people which speak the language Afaan Oromoo as their second language. In Somalia there are also about some speakers of the language. According to (Tilahun, 2002) in Kenya, the Ethnologue also lists 322,000 speakers of Borana and Orma, two languages closely related to Ethiopian Oromo. Currently, the above figure increased more than the above figures indicated.

In addition to this, in Africa, it is the language with the third (3rd) most speakers, after Arabic and, Swahili.

Besides native speakers, a number of members of other ethnicities who are in contact with the Oromoos speak Afaan Oromoo as a second language. For instance, the Omotic-speaking Bambassi and the Nilo-Saharan-speaking Kwama in north-western Oromiya are the notables (Asafa, 2005). The Oromoos adopted the Latin alphabet for their writing system which replaced the old geez alphabet. In the next section, the Afaan Oromoo language emphasising on the alphabet is discussed in detail.

3.2. Afaan Oromoo Alphabets (Sagaleewani fi Loqoda)

Afaan Oromo language is one of the Cushitic languages (such as Somali, Afar, Sidama, Geedo, and ancient Egyptian), the language of the ancient Cush that was spoken all over East Africa (Roba, 2002). According to (Roba, 2002) Afaan Oromoo is the language predominantly spoken in every corner of the country.

Afaan Oromo is a phonetic language, which means that it is spoken in the way it is written. Afaan Oromo uses the Roman alphabet but it has its own consonants and vowels.

Afaan Oromoo has 28 letters called 'qubee'. Actually, a new letter 'Z' is now included in the alphabets as there are words which require letter 'Z'. Therefore, now it is correct to say that Afaan Oromo has 29 qubee. The Afaan Oromoo qubee has both capital (qubee gurgudda) and small letters (qubee xixiqaa). The Afaan Oromoo qubee with their respective pronunciation is presented in the following table 3.1.

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

Table 3.1: Afaan Oromoo alphabets and there pronunciation (extracted from: IPA)

The effects of the vowels and consonants on the production of speech are found to be of great deal. In the following section the characteristics and the effects of vowels (dubbachiftuu) and consonants (dubbifama) are seen in detail.

Afaan Oromoo Vowels: (Dubbachiftuu)

Afaan Oromo vowels are represented by the five letters, **a, e, o, u and i.** All vowels are pronounced basically the same way throughout Oromiya. These vowels when stressed may be opened: deem*u* (go), nyaadh*u* (eat) or closed: b*a*da, r*a*fi. The following are the Afaan Oromoo vowels with manner of the vowels created.

Vowels			
	Front	Central	Back
Close	i /ɪ/, ii /iː/		u /ʊ/, uu /uː/
Mid	e /ɛ/, ee /eː/		o /ɔ/, oo /oː/
Open		a /ʌ/	aa /ɑː/

Table 3.2. Afaan Oromoo vowels (adapted from: IPA)

The Afaan Oromo vowels always are pronounced in sharp and clear fashion which means each and every word is pronounced strongly, for example:

Vowels	Examples
A	Ar'bba, Farda, Haadha
E	Gannaale, Waabee, Noole, Roobale, colle
I	Arsii, laali, Rafi, Lakki, Sirbbi
O	Oromo, Cilaalo, Haroo, caancco, Danbidoollo
U	Ulfaadhu, Guddadhu, dubadhuu, arbba guugu, Ituu

Table 3.3. Afaan Oromoo vowels with examples (extracted from: Taha Roba, 2002)

Afaan Oromoo Consonants -Sagaleewwan(Dubifamtoota)

Afaan Oromoo consonants also show a systematic of opposition of voiceless, aspirated, voiced and voiceless ejective stops and affricates. The voiceless uvular stop has no ejective or voiced counterparts. Most Afaan Oromo constants are uttered in clear fashion even though some exceptions and few special combinations. Some examples of those consonants are;

A. The consonant "g" has a hard sound. Gaari, gadi bayi, gargaari.

B. The combinations NY and DH have a hard sound. e.g Nyaadhu, Dhugi.

Afaan Oromoo Double Consonants - Sagalewwan Dachaa

All Afaan Oromo consonants except the combination consonants (quubeewwaan dachaa) **ch**, **dh**, **ph**, **ny** and **sh** have double consonant combinations if the syllable is stressed. Failure to make this distinction results in miscommunication. Examples:

Bilisumma, adda, malamaltumma, fardda, lolttu. In addition to this Afaan Oromoo consonants have the following characteristics which is summarized in table3.4

Consonants (dubbifamaa)						
		Bilabial (Labiodental)	Alveolar (Retroflex)	Palato-alveolar (palatal)	Velar	Glottal
Stops and Fricatives	Voiceless	(p)	T	ch /ʧ/	K	' /ʔ/
	Voiced	B	D	j /ʤ/	g	
	Ejective	ph /p'/	x /t'/	c /ʧ'/	q /k'/	
	Implosive		dh /ɗ/			
Fricatives	Voiceless	F	S	sh /ʃ/		H
	Voiced	(V)	(z)			
Nasal		M	N	ny /ɲ/		
Approximants		W	I	y /j/		
Flap/Trill			R			

Table 3.4. Afaan Oromoo consonants (Dubbifamaa) (extracted from: IPA)

Stress in Afaan Oromoo - Jabaataa

Some Afaan Oromo words are pronounced with the stress on the last syllable: for example: malamaltumma, laggen, gaarre.

On the other hand, few words are stressed on the first syllable. These words always have a combination consonant: e.g nyaadhu, dhayi, nyaara nyaapha (foreigner).

Therefore Afaan Oromoo language recognizer developed is in line with all those characteristics of the Afaan Oromoo language. Some other considerable factor was also considered so as to make the recognizer model better represent the language under study. Another important factor we need to consider is the sounds and orthography of the language Afaan Oromoo which is described in the following section.

3.3. Sounds and orthography

Consonant and vowel phonemes

Like most other Ethiopian languages, whether Semitic, Cushitic, or Omotic, Afaan Oromo has a set of ejective consonants, that is, voiceless stops or affricates that are accompanied by glottalization and an explosive burst of air. Afaan Oromoo has another glottalized phone that is more unusual, an implosive retroflex stop, "dh" in Afaan Oromo orthography, a sound that is like an English "d" produced with the tongue curled back slightly and with the air drawn in so that a glottal stop is heard before the following vowel begins(Bulti, 2003).

Afaan Oromoo has the typical Southern Cushitic set of five short and five long vowels, indicated in the orthography by doubling the five vowel letters. The difference in length is contrastive, for example, *hara* 'lake', *haaraa* 'new'. Gemination is also significant in Afaan Oromoo. That is, consonant length can distinguish words from one another, for example, *badaa* 'bad', *baddaa* 'highland'.

In the Qubee alphabet, a single "letter" consists either of a single symbol or a digraph ("ch", "dh", "ny", "ph", "sh"). Gemination is not obligatorily marked for the digraphs,

though some writers indicate it by doubling the first symbol: *qopphaa'uu* 'be prepared'.

In the following section, special Afaan Oromoo characters to be considered as of the IPA symbol for a phoneme is shown in brackets where it differs from the Afaan Oromoo letter. The phonemes /p v z/ appear in parentheses because they are only found in recent loan words.

It is all notable that, there have been minor changes in the orthography since it was first adopted: <x> ([t’]) was originally represented as "th", and there has been some confusion among authors in the use of "c" and "ch" in representing the phonemes /tʃ’/ and /tʃ/, with some early works using "c" for /tʃ/ and "ch" for /tʃ’/ and even "c" for different phonemes depending on where it appears in a word. This article uses "c" consistently for /tʃ’/ and "ch" for /tʃ/ (Bulti, 2003).

Like most other Afro-Asiatic language, Afaan Oromoo has two grammatical genders, masculine and feminine, and all nouns belong to either one or the other. Grammatical gender in Afaan Oromo enters into the grammar in the following ways:

- Verbs (except for the copula *be*) agree with their subjects in gender when the subject is third person singular (*he* or *she*).
- Third person singular personal pronouns (*he*, *she*, *it*, etc., in English) have the gender of the noun they refer to.
- Adjectives agree with the nouns they modify in gender.
- Some possessive adjectives ("my", "your") agree with the nouns they modify in some dialects.

Except in some southern dialects, there is nothing in the form of most nouns that indicates their gender. A small number of nouns pairs for people, however, end in -*eessa* (m.) and -*eettii* (f.), as do adjectives when they are used as nouns: *obboleessa* 'brother', *obboleettii* 'sister', *dureessa* 'the rich one (m.)', *hiyyeettii* 'the poor one (f.)'.

Grammatical gender normally agrees with biological gender for people and animals; thus nouns such as *abbaa* 'father', *ilma* 'son', and *sangaa* 'ox' are masculine, while nouns such as *haadha* 'mother' and *intala* 'girl, daughter' are feminine. However, most names for animals do not specify biological gender.

The words phrases and simple sentence selected for this researcher are selected and collected considering all the above characteristics of the Afaan Oromoo language and the phonemes. This helped the researcher in finding the characteristics of the language and representing the language for better results of the thesis. Furthermore, the expert in the domain (linguistics and phonetics) was consulted from the linguistics and phonetics of Afaan Oromoo.

Afaan Oromoo Nouns

Afaan Oromoo nouns are presented in their respective examples of each type so as to simplify for the nouns used in the language under investigation.

Nominative

The nominative is used for nouns that are the subjects of clauses.

- *Ibsaa* man's name, *Ibsaan* 'Ibsaa (nom.)', *konkoolaataa*, *qaba* 'he has', *Ibsaan makiinaa qaba* 'Ibsaa has a car'

Genitive

The genitive is used for possession or "belonging"; it corresponds roughly to English *of* or -'*s*. The genitive is usually formed by lengthening a final short vowel, by adding *-ii* to a final consonant, and by leaving a final long vowel unchanged. The possessor noun follows the possessed noun in a genitive phrase. Many such phrases with specific technical meanings have been added to the Oromo lexicon in recent years.

- *obboleetti* 'sister', *namicha* 'the man', *obboleetti namichaa* 'the man's sister'
- *hojii* 'job', *Caaltuu*, woman's name, *hojii Caaltuu*, 'Caaltuu's job'
- *barumsa* 'field of study', *afaan* 'mouth, language', *barumsa afaanii* 'linguistics'

In place of the genitive it is also possible to use the relative marker *kan* (m.) / *tan* (f.) preceding the possessor.

- *obboleetti kan namicha* 'the man's sister'

Dative

The dative is used for nouns that represent the recipient (*to*) or the benefactor (*for*) of an event. The dative form of a verb infinitive (which acts like a noun in Oromo) indicates purpose. The dative takes one of the following forms:

- Lengthening of a final short vowel (ambiguously also signifying the genitive)

- *namicha* 'the man', *namichaa* 'to the man, of the man'

Instrumental

The instrumental is used for nouns that represent the instrument ("with"), the means ("by"), the agent ("by"), the reason, or the time of an event. The formation of the instrumental parallels that of the dative to some extent:

- *-n* following a long vowel or a lengthened short vowel; *-iin* following a consonant

- *harka* 'hand', *harkaan* 'by hand, with a hand'

- *Afaan Oromoo* 'Oromo (language)', *Afaan Oromootiin* 'in Oromo'

Locative

The locative is used for nouns that represent general locations of events or states, roughly *at*. For more specific locations, Oromo uses prepositions or postpositions. Postpositions may also take the locative suffix. The locative also seems to overlap somewhat with the instrumental, sometimes having a temporal function. The locative is formed with the suffix *-tti*.

- *guyyaa* 'day', *guyyaatti* 'per day'
- *jala, jalatti* 'under'

Ablative

The ablative is to represent the source of an event; it corresponds closely to English *from*. The ablative, applied to postpositions and locative adverbs as well as nouns proper, is formed in the following ways:

keessa 'inside, in', *keessaa* 'from inside'

- *finfinnee* 'finfine(Addis Ababa)', *Finfinneedhaa* 'from Finfinnee (Addis Ababa)'
- *gabaa* 'market', *gabaadhaa* 'from market'

Hence the words, phrases and sentences from Afaan Oromoo language selected constituting the verbs and nouns which fulfil the above characteristics of the language whenever necessary to achieve the objective of the research conducted.

To put in to effects the recognizer of the Afaan Oromoo language, in the next chapter the tools, techniques and methodologies are selected and discussed for this research in detail.

CHAPTER FOUR

METHODOLOGY

4.1. Introduction

In this chapter, the tools, techniques and procedures used for this research is discussed in detail. Accordingly, in this research the HMM approach with statistical method were used. Therefore this chapter presents the tools, techniques and approaches used in this specific research.

4.2. Definition of HMM

According to Rabiner L. R. and B. H. Juang (Rabiner & Juang 1993), a hidden Markov model is a doubly embedded stochastic process with an underlying stochastic process that is not directly observable (it is hidden) but can be observed only through another set of stochastic processes that produce the sequence of observations.

HMMs use a process to model the changing statistical characteristics that are only probabilistically manifested through actual observations. The state sequence is *hidden*, and can only be observed through another set of observable stochastic processes. Each hidden state of the model (or the transition between states) is associated with a set of output probability distribution or continuous Probability Density Functions (PDF). The mask between hidden state sequences and the observable stochastic process is characterized by the output probabilities.

Accordingly we do have n states $s_1 ... s_n$ *in an HMM, and the states are connected* and in addition to this *t*he output symbols are produced by the states or edges in HMM. Hence an observation $O = (o_1 ... o_T)$ is a sequence of output symbols, given an observation; we want to recover the hidden state sequence. The following diagram figure 4.1 depicts the hypothetical example for five states HMM that undergoes a change of states according to a set of probabilities associated with the states.

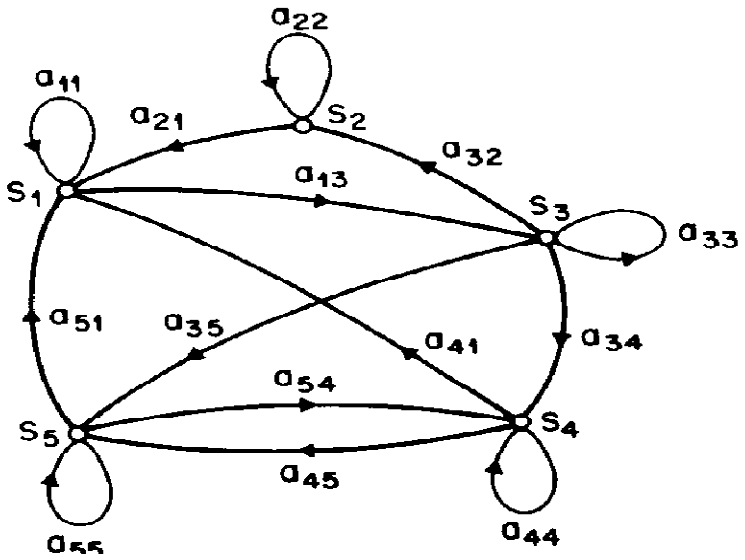

Figure 4.1 A Markov chain with five states (extracted from: Rabiner, 1989)

4.4. Elements of HMM

This part illustrates the elements involved in HMM and how formally these elements generate the observation sequences.

N- Number of hidden states in the model: although the states are hidden, for many practical applications there are some physical significance attached to the states or to the set of the states of the model. In general the states are interconnected in such a

way that any state can be reached from any other state. Here we denote the individual states as;

$$S = \{s_1, s_2, ..., s_N\} \quad \text{.....4.1}$$

M- The number of distinct observation symbols per states: the observation symbols correspond to the physical output of the system being modelled. Therefore, we denote the individual symbols as;

$$V = \{v_1, v_2, ..., v_M\} \quad \text{.....4.2}$$

The state-transition probability distribution $A = \{a_{ij}\}$ where,

$$a_{ij} = P(q_{t+1} = j \mid q_t = i) 1 \le i, j, \le N \quad \text{.....4.3}$$

In the above example, any state can reach to any other state in a single step.

Observation symbol probability distribution in state j, $B = \{b_j(k)\}$, where;

$$b_j(k) = P(v_k t \mid q_t = S_j\} \quad \text{.....4.4}$$

For $1 \le j \le N$ and $1 \le k \le M$

The initial state distribution $\pi = \{\pi_i\}$ where;

$$\pi_i = P(q_1 = S_i) \quad \text{.....4.5}$$

For $1 \le i \le N$

Given the appropriate values N, M, A, B, and π the HMM can be used as a generator to a given sequence of observations.

$O = O_1, O_2, ..., O_T$ 4.6

Where each of the observation O_t is one of the symbols from V_t and T is the number of observations in the sequence as follows;

a. choose an initial state $q_t = S_i$ according to the initial state distribution π

b. set t=1

c. choose $O_t = v_k$ according to the symbol distribution in state S_i that is $b_j(k)$.

d. Transit to new state $q_{t+1} = S_j$ according to the state transition probability distribution for state S_i, that is, a_{ij}.

e. Set $t = t + 1$ and return to step c, if $t < T$; otherwise terminate the procedure

It can be seen from the above discussion that a complete specification of the HMM requires, specification of two model parameters (N and M), specification of observation symbol and the specification of the three probability measure A, B and π. Therefore the entire model (λ) is represented as;

$\lambda = (A, B, \pi)$ 4.7

Basic Assumptions of HMM

While using HMM we need to consider three basic assumptions (Rabiner, 1989). These assumptions are made for mathematical and computational traceability.

a. The Markov Assumption

The first assumption states that history has no influence on the chain's future evolution if the present is specified. 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 (Rabiner and Juwang, 1993).

Therefore the above assumption can be mathematically expressed as;

$$a_{ij} = P(q_{t+1} = j \mid q_t = i) \quad \text{.....4.8}$$

where a_{ij} is the transition probability from state i to j; q_{t+1} is the next state after the current state q_t.

However generally the next state may depend on past k states and it is possible to obtain such a model, called a K^{th} order HMM by defining the transition probabilities as follows;

$$a_{i1i2\ldots ikj} = P(q_{t+1} = j \mid q_t = i_1, q_{t-1} = i_2, \ldots q_{t-k+t} = i_k) \quad \text{.....4.9}$$

Where $1 \le i_1, i_2, \ldots i_k, j \le N$

But it is seen that a higher order HMM will have a higher complexity (Warakagoda and Høgskole, 1996). Even though the first order HMMs are the most common, some attempts have been made to use the higher order HMMs for speech recognition and other related stochastic process modelling.

b. The Stationary Assumption

The assumption here is the stationary assumption of HMMs. It states that state transition probabilities are independent of the actual time at which the transitions take place (Rabiner, 1989). This assumption confirms that the transitions are independent of the point of time at which the transition took place.

Therefore mathematically for any given time t_1 and t_2 we can show as;

$$(q_{t_1+1} = j \mid q_{t_1} = i) = (q_{t_2+1} = j \mid q_{t_2} = i) \quad \text{.....4.10}$$

Where q_{t_1+1} is the next state at t_1 and q_{t_2+1} is the next state at t_2 for the state variables i and j.

c. The Output Independence Assumption

The output independence hypothesis states that neither chain evolution nor past observations influence the present observation if the last chain transition is specified (Rabiner, 1989). In other words, the current observations or outputs are statistically independent of the previous observations or outputs.

By considering the following sequence of observations, the fore mentioned assumption can be mathematically explicated as;

$$O = o_1, o_2, ..., o_T \quad \text{.....4.11}$$

Thus for a given HMM λ;

$$P(O \mid q_1, q_2, ...q_T, \lambda) = \prod_{t=1}^{T} P(o_t \mid q_t, \lambda) \quad \text{.....4.12}$$

But unlike the other two assumptions of HMMs, the output independence assumption has limited validity (Warakagoda and Høgskole, 1996) and sometimes considered as a sever weakness of HMM. In the next part the three problems and their respective solution are seen so as elaborated how it works.

Types of HMM

Generally there are three types of HMM models available.

Continuous HMMs

Quantization errors can be eliminated by using a continuous density model, instead of VQ codebooks. In this approach, the probability distribution over acoustic space is modelled directly by a continuous Probability Density Function (PDF), which is typically a mixture of Gaussian functions.

For example, when observed data, x, are n-dimensional vectors and the Gaussian mixture density is used, the probability of generating data x given the transition from state i to state j, $\Pr(X \mid ij)$, becomes

$$pr(x \mid ij) = \sum_{m=1}^{M} w_{ij}^{(m)} N_{ij}^{(m)}(x) \quad \text{.....4.13}$$

Where N(m) is the $m-th$ guassian w(m) the mixture weight and $\sum_m w_{ij}^{(m)}$ to make $f(x)pr(x \mid ij)dx = 1$. Suppose there are n CHMMs in the system and each HMM has a arcs. The number of parameters is then na for transition probabilities, plus naM for the densities, assuming that there is one initial and one final state for all HMMs. In practice the number of mixtures, M , is something in between ten thirty.

Discrete HMMs

In this approach, the entire acoustic space is divided into moderate number of regions, by a clustering procedure known as VQ. The centroid of each cluster is represented by a scalar codeword, which is an index into a codebook that identifies the corresponding acoustic vectors. Each input frame is converted to a codeword by finding the nearest vector in the codebook.

The HMM output symbols are also codewords. Thus, the observation probability distribution over acoustic space is represented by a simple look-up table over the codebook entries.

DHMMs do not make any assumption about the form of the output distribution. Therefore, they can well represent all possible distributions. Moreover, computing the output probability $pr(x \mid ij)$ becomes only a table lookup, rather than the intensive multiplications and additions required for CHMMs. That is,

$$pr(x \| ij) = b_{ij}(k) \quad \text{.....4.14}$$

where k is the vector quantized symbol for x.

However, there are two disadvantages. Here, many more parameters (and thus more training data) are necessary for DHMMs than for CHMMs. With DHMMs, the number of parameters is na plus naL, where n is the number of models in a system, a is the number arcs per HMM, and L is the size of the VQ codebook. Normally L is about 256 versus M = 20 in CHMMs.

Semi-Continuous HMMs

To relieve the VQ distortion in DHMMs, semi-continuous HMMs (SCHMMs) or tiedmixture HMMs were proposed independently. When SCHMMs are used, the discrete output distribution $b_{ij}(k)$ is retained as the distribution representation. In addition, each VQ codeword k is modelled by a density function $f_x(k)$, which can be explained as the likelihood of vector x belonging to codeword k. The simple table lookup $b_{ij}(k)$ in DHMMs to find the probability of generating an acoustic vector x given the transition from state i to state j is then replaced by;

$$\Pr(x \mid ij) = \sum_{k=1}^{L} b_{ij}(k)\, f_k(x) \quad \text{.....4.15}$$

where L is the VQ level, that is, the size of the VQ codebook. To save computation, SPHINX-Il usually uses only the top 4 best-matched VQ prototypes to represent the input vector, rather than using all the L codeword's. Each f_k is assumed to be a

single Gaussian with a diagonal covariance in SPHINX. Moreover, we found that normalizing the $f_k(x)$'s such that increased the stability of our training routines. This is because the density values can be arbitrarily small or large, which may cause floating-point exceptions, especially when a double precision floating-point representation is not used. The floating-point exception problem becomes serious when multiple independent features are used (H.Doe, 1998).

Compared with CHMMs which require the computation of *naM* densities for every piece of data x, SCHMMs require only the computation of L densities since all transitions are tied with the same mixtures. With much less computation, SCHMMs perform at least as well as CHMMs in terms of recognition accuracy (H.Doe, 1998).

The Basic Problems of HMM

According to Rabiner (1989), given the above fact and discussions of the HMM elements and basic assumptions we have three basic problems of HMM model.

1. Given observations $O = (o_1, o_2, ...o_T)$ and model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(O \mid \lambda)$, the probability of the observation sequences given that the model. Because:
    - Here hidden states complicate the evaluation
    - Given two model λ_1 and λ_2, this can be used to chose the better one.

This problem refers to the evaluation problem given the model and sequences of observations, how do we compute the probability that the observed sequences was

produced by the model. Furthermore we can also view the problem as one of the scoring how well the given model matches a given observation sequence.

2. Given the observation sequence $O = (o_1, o_2, ..., o_T)$ and model λ how we choose a corresponding state sequence $q = (q_1, q_2, ..., q_T)$ which is optimal in some meaningful sense or best explains the observations.
    - Optimality criterion has to be decided (e.g. maximum likelihood)
    - Explanation for the data.

Here the problem is all about uncovering the hidden parts of the model or finding the correct state sequences. In case of practical situation, it is important to note that there is correct matches rather it is searching for optimality criteria to solve those problems. A typical uses might be to learn the structure of the model, to find optimal state sequence for continuous speech recognition.

3. Given $O = (o_1, o_2, ...o_T)$, estimate model parameters $\lambda = (A, B, \pi)$ how do we adjust the model parameter that maximizes $P(O \mid \lambda)$.

This problem refers to the learning problem of HMM.

Solution to the Three Problems of HMM

Solution to problem 1: Evaluation and forward algorithm

To calculate the probability of the observation sequence, given the model, one of the straight forward procedures is enumerating every possible state sequence of length T. Recursive procedures like forward and Backward Procedures exist which can compute $P(O \mid \lambda)$. Here the procedure to perform this is given below for both, forward and backward;

Forward Procedure

Initialization

$$\alpha_1(i) = \pi_t b_t(o_1) \text{ where} 1 \le i \le N$$

Induction

$$\alpha_{t+1}(j) = \left[\alpha_t(i) a_{ij}\right] b_j(o_{t+1}) \text{ where } 1 \le t \le T-1, 1 \le j \le N$$

Termination

$$P(O \mid \lambda) = \sum_{t=1}^{N} \alpha_T(i)$$

Forward variable $\alpha_t(i)$ is defined as $\alpha_t(i) = P(o_1 o_2 o_3 ..., q_i = i \mid \lambda)$ that is the probability of the potential observation sequence until time t and state i at time t, given the model λ.

Backward Procedure

Initialization

$\beta_T(i) = 1$ where $1 \le i \le N$

Induction

$$\beta_T(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \text{ where } T-1 \le t \le 1, 1 \le i \le N$$

Solution to problem 2: Decoding and Viterbi algorithm

Finding the optimal sequence associated with a given observation. Viterbi algorithm finds the single best sequence q for the given observation sequence O. The following equations are presented which is the Viterbi algorithm. Viterbi algorithm returns the best possible observation sequence $q*$ and also probability score for that state sequence $p*$. This is mathematically given as;

Initialization

$\delta_1(i) = \pi_i b_i(o_1)$ where $1 \le i \le N$

$\Psi_1(i) = 0$

Recursion

$$\delta_t(j) = \max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij} \right] b_j(o_t) \text{ where } 2 \le t \le T$$

$$\Psi_t(j) = \arg\max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij} \right] b_j(o_t) \quad \text{Where} \quad 1 \le j \le N$$

Termination

$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

$$q_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)]$$

Path backtracking

$q_t^* = \psi_{t+1}(q_{t+1}^*$ where $t = T-1, T-2, T-3, T-4, \ldots 1$

Solution of problem 3: Learning and Baum-Welch algorithm

The third problem is the most difficult problem of the three, problem of adjusting the model parameter λ = (A,B,π) such that the probability of P(O | λ) maximized. One of the popular methods used for locally maximization of the probability of the observation using an iterative procedure such id Baum - Welch (Rabiner, 1989) method. But, the Baum-Welch re-estimation procedure also suffers from the following problems;

- Has numerical problems and hence hard to implement
- Needs special scaling
- Needs multiple observation sequences

In order to describe the procedures for re-estimation (iterative update and improvement) of HMM parameters, we first define $\xi_t(i, j)$, the probability of being in state S_i at time t and state S_j at time $t+1$, given the model and observation sequence;

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j \mid O, \lambda$$

From the definitions of the forward and backward variables, that we can write $\xi_t(i,j)$ in the form;

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)}$$

$$= \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}$$

where the numerator term is just $P(q_t = S_i, q_{t+1} = S_j, O \mid \lambda)$ and the division by $P(O \mid \lambda)$ gives the desired probability measure.

The posterior probability of being in state S_i at time t given the observation sequence and the model as;

$$\delta_t(i) = P(q_t = S_i \mid O, \lambda)$$

Having the values of δ and ξ one can define the update rules as;

$$\bar{\pi}_i = \delta_i(1)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \delta_i(t)} \text{ and } b_{ij}(\bar{k}) = \frac{\sum_{t=1}^{T} \delta_{O_t, O_k \gamma_i(t)}}{\sum_{t=1}^{T} \delta_i(t)}$$

Here the main objective is learning from the training data.

Language Model (N-gram)

A language model is used in speech recognition systems and automatic translation systems to improve the performance of such systems (Strauss, 1992). The language model for speech recognition is one of the important aspects to consider here. Therefore one the dominant language model used in line with the HMM approach, the N-gram language model, introduced below.

An N-Gram grammar is a representation of an N-th order Markov language model in which the probability of occurrence of a symbol is conditioned upon the prior occurrence of N-1 other symbols (Brown, 2001).

N-Gram grammars are typically constructed from statistics obtained from a large corpus of text using the co-occurrences of words in the corpus to determine word sequence probabilities.

N-Gram language models are traditionally used in large vocabulary speech recognition systems to provide the recognizer with an a-priori likelihood $P(W)$ of a given word sequence W . The N-Gram language model is usually derived from large training texts that share the same language characteristics as expected input. This information complements the acoustic model $P(W \mid O)$ that models the articulatory features of the speakers. Together, these two components allow a system to compute the most likely input sequence in the such a way that;

$\mathbf{W' = \text{argmax}_W P(W \mid O)}$,

where **O** is the input signal observations as **$W' = \text{argmax}_W P(O|W) P(W)$.**

The methodology discussed above is well accepted and most of the researchers were using this because of all the above facts that the technique equipped. Therefore the researcher intended to use this methodology in line with its components discussed so far.

In addition to this the next part of this sub topic elaborates about the tools, sphinx system, how the techniques selected and used in this experiment. To this end the following section also elaborates the selected tool over the others.

Tools Used For Speech Recognition

According to different literature, there are various tools used in automatic speech recognition. Among others, HMM toolkit (HTK) and SPHINX systems are predominantly used in various researches because of the popularity and robustness that the tools provide for the researchers. Beyond this, there are also different yardsticks that the evaluators use to select their tools. Up on consulting all those literatures the researcher selected the appropriate tool for Afaan Oromoo Continuous Speech Recognizer which is discussed here.

Sphinx system

SPHINX system has been developed at Carnegie Mellon University. Currently there are SPHINX 2, 3, 3.5 and 4 decoder versions and SphinxTrain used for training purposes.

The system has been developed entirely by java and it is platform independent beyond the accuracy level it provides. Furthermore as the sphinx system has both versions for training and decoding, the following section discussed the two components separately.

Sphinx Train

Sphinx train is module found for training of the speech corpus mainly containing the dictionary and language model, the model and model loader and the frontend. So as to use the sphinx train we need to download freely from CMU home and use with changes of the above requirements. SphinxTrain performs triphone tying by constructing decision trees; however no phone classification file is needed. Questions are somehow formed but this particular mechanism is somehow obscured. Instead of setting stoppage condition for further splitting, the number of tied states is manually set up by the designer. Furthermore only states can be tied (not mean, variances, etc.).

Apart of SphinxTrainer there is CMU statistical modeling tool that can be used to construct words counts, bigrams and trigrams counts, various backoff bigram and trigram language models, perplexity, out of vocabulary ratios, etc. SphinxTrain can further performs semi-continues model training with vector quantization.

Sphin4 Decoding Model

Sphinx4 is a state-of-art speech recognition system written entirely in JavaTM programming language. It was created by joint collaboration between Sphinx group at CMU, Sun Microsystem Laboratories, Mitsubishi Electric Research Labs and

Hewelett Packard, with contribution from the University of California at Santa Cruz and the Massachusetts Institute of Technology.

Sphinx4 started out as a port of sphinx3 to the Java programming language, but involved in to a recognizer designed to be much more flexible than sphinx3, thus becoming an excellent platform for speech research.

Sphinx-4 is a modular and pluggable framework that incorporates design patterns from existing systems, with sufficient flexibility to support emerging areas of research interest. The framework is modular in that it comprises separable components dedicated to specific tasks, and it is pluggable in that modules can be easily replaced at runtime. To exercise the framework, and to provide researchers with a working system, Sphinx-4 also includes a variety of modules that implement state-of-the-art speech recognition techniques.

Generally Sphinx-4 is a flexible, modular and pluggable framework to help foster new innovations in the core research of hidden Markov model (HMM) recognition systems.

Therefore to support above feature of the sphinx4 model in the following section we are going to view the model architecture with its components so as have clear understanding on the components of the decoding model.

Sphinx4 Architecture

In this section we describe the various components of Sphinx-4, and how they work together during the recognition process. First of all, let's look at the architecture diagram of Sphinx-4.

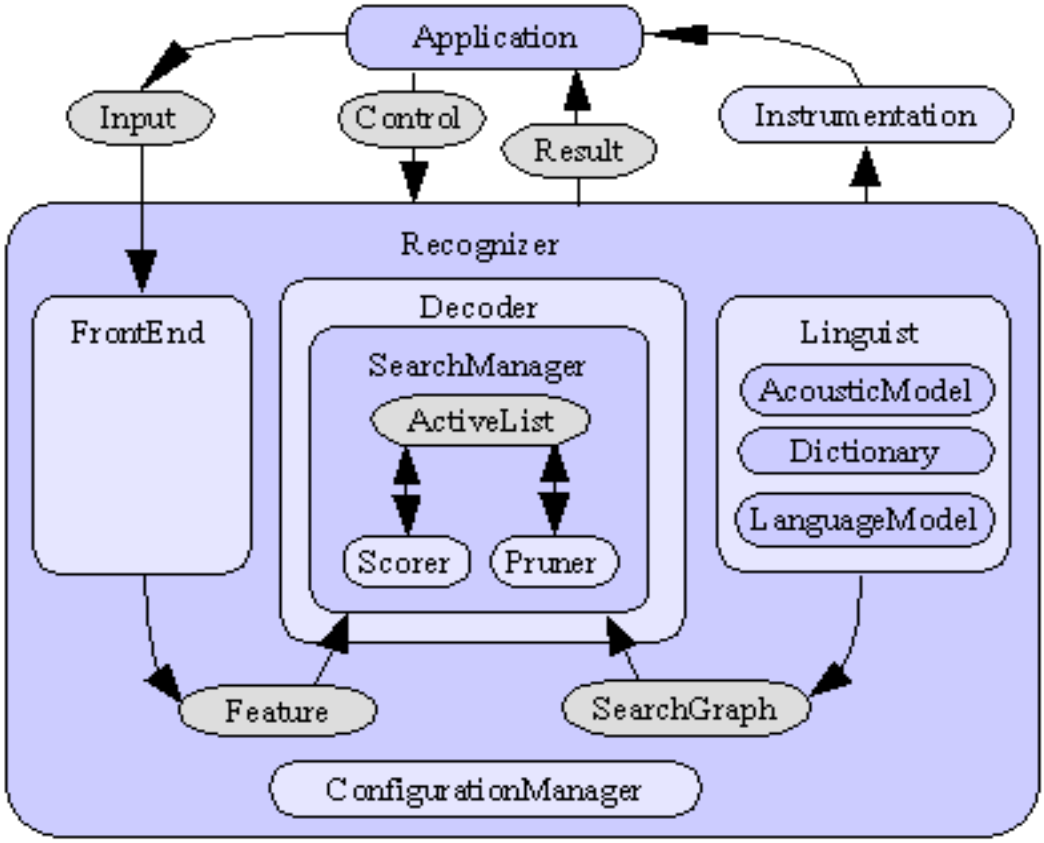

Figure4.2. Sphinx4 Architecture

The recognizer constructs the front end (which generates features from speech), the decoder, and the linguist (which generates the search graph) according to the configuration specified by the user. These components will in turn construct their own subcomponents.

For instance, the linguist will construct the acoustic model, the dictionary, and the language model. It will use the knowledge from these three components to construct a search graph that is appropriate for the task. The decoder will construct the search manager, which in turn constructs the scorer, the pruner, and the active list.

Most of these components represent interfaces. The search manager, linguist, acoustic model, dictionary, language model, active list, scorer, pruner, and search graph are all Java interfaces. There can be different implementations of these interfaces. The implementation to be used is specified by the user via the configuration file, an XMLbased file that is loaded by the configuration manager. In the configuration file, a user can also specify the properties of the implementations.

Generally the sphinx4 decoder has three basic components, each with their specific tasks, discussed here.

Front End: Accepts audio data in the form of audio files (.wav, batch files) as well as live voice from a microphone. The front end takes the input audio signal and puts it through an analysis process that includes pre-emphasis, signal segmentation, and frequency analysis.

Linguist: Constructs a search graph using the knowledge base. The knowledge base holds all the information for matching the incoming voice data with actual words and phrases. It is made up of three parts:

1. Acoustic Model: Holds sound data packets. Each sound packet represents a part of a sound that we make when we say a word. Each data packet represents a

portion of a syllable in a word. A few data packets will make up one syllable and many data packets will eventually make up a word. This database is large because of the extent of the English language and the different portions that can make up each word.

2. Dictionary: Also called the lexicon, this contains all the words that the speech recognition engine will recognize and how they are pronounced.

3. Language Model: Contains information about probabilities of outcomes for single words and phrases. Given a set of sound data packets, it will help determine which word that it is likely to be. It also holds grammar rules so that it can determine what word will likely follow another word in a phrase. This is called an N-gram method of recognition. For example, a tri-gram would look at the previous two words of a phrase and then using probability to determine the next likely word to come.

Decoder: Brings together the Front-End and the Knowledge Base to output the result to the user.

Having studied all those important methodologies and tools for the Afaan Oromoo Continuous speech recognizer, the appropriate methodologies, tools and algorithms are indicated within the experimentation and the next chapter shows as the design and implementation of the recognizer.

CHAPTER FIVE

EXPERMENTATION

In this chapter, the investigation of experiments for Continuous Afaan Oromoo Speech Recognizer along with their discussions is presented. As the ultimate goal of any speech recognition prototype is checking the suitability and applicability of the tools and approaches for the language under investigation, prototype developed for Continuous Afaan Oromoo Speech Recognizer which is elaborated here.

This experiment is performed on the machine, Toshiba satellite laptop with RAM 2.0 GB, Processor 2.1GHz Dual processor and Hard disk capacity of 160GB, with headset microphone, sound card 32-bit and windows vista operating system.

The model constructed is also viewed from two perspectives, namely, Context dependent and context independent, by building two language models. Defining context independent models is directly related to phoneme distance measures and triphone based. Whereas, the context dependent takes the entire words, phrases and sentences for the dictionary and other requirements. The two models are the monophone model which directly takes the words, phrases and sentences of the dictionary and the triphone model that takes the phonemes generated so as to perform the tasks.

5.1. Continuous Afaan Oromoo speech recognizer design

The adapted recognizer design used in the entire development of the prototype is given in Figure 5.1.

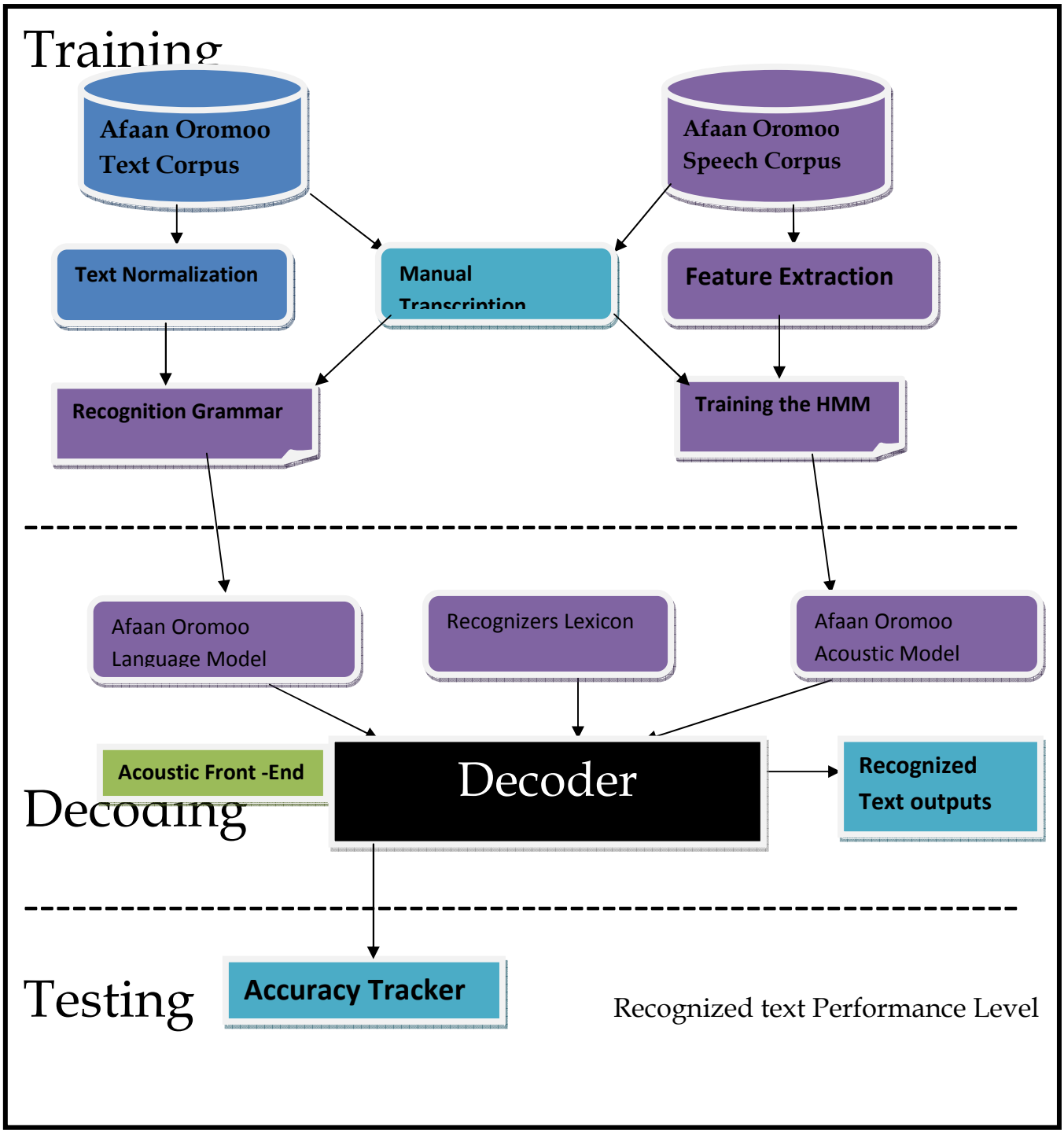

Figure 5.1. Afaan Oromoo continuous speech recognizer design

5.2. Data Recording and Pre-processing

For this research, the data collection was performed in Ziway town, which is at a distance of 160 KM from the capital Addis Ababa to the south East. Accordingly the selected Afaan Oromoo words, phrases and simple sentences are recorded in Batu elementary and Secondary School from 30 selected students and peoples of which 15 are females and the remaining 15 are males. This is mainly because the researcher couldn't get access to sound proofed rooms in Addis Ababa and the speech recording task requires silent environment and the noise might reduce the performance of the recognizer.

The content of the recorded corpus is similar to the previously prepared corpus for this purpose and speakers uttered all the 70 phrases by uttering each of the seven (7) phrases at ones. Hence for one speaker we have recorded ten (10) wav files. So as to make the utterances quit similar the researcher also used sample utterance for the speaker as some speaker need training of how to utter the phrases and this also performed by the researcher.

Here, we also used some selected age group for uttering the phrases as age has its own impact on the speech recognition tasks. The different age group selected for this researcher is indicated in the following table 5.1

Age Range	Number of speaker
0-15	5
16-30	20
31-45	5

Table 5.1. Age category of the selected speakers

The recorded speech data set or the whole speech corpus is classified in to 2/3 by 1/3 for training and testing respectively. After the data are recorded, using the Praat software for speech recoding and labeling, the necessary preprocessing was performed for Afaan Oromoo continuous speech recognition. The recoded speech data are assigned with specific identification number for users followed by their sex and the phrase uttered number. The following table 5.2 indicates the proportion of the Afaan Oromoo corpus used for experiment as classified for training and testing the model respectively.

Data set	**Number of speakers**	**Number of phrases per speaker**
Training	20	70
Testing	10	70

Table 5.2. The corpus for Afaan Oromoo speech

The recoded data are mono and with 16 KHz and saved as wav files as the sphinx train supports the wav file formats. These files are recorded using PRAAT which is open source software. For instance, for the first speaker, the audio wav file identification numbers ranging from AO001F01 to AO001F10 was assigned. The assignment the file names are performed for all speakers in similar fashion. From research ethical view point, this helps to keep confidentiality and anonymity of the speakers.

5.2.1. The phoneme sets extraction

Phoneme sets extraction is one of the important task performed while developing the model. The phone sets are extracted from the dictionary prepared manually and the knowledge base of the CMU which produces the dictionary for a given corpus with

some modification manually as the knowledge base is specifically designed for English language that couldn't support the characteristics of the language Afaan Oromoo.

From the dictionaries prepared unique phones from each dictionary for respective phrases were performed and the unique phones of the dictionaries are arranged alphabetically. The SIL phoneme is added to the phone list so as to handle the silence filler dictionary that might cause errors during training.

For the given corpus, constituting the phrases and simple sentence, the dictionaries were prepared and generated a phoneme in the following manner. For instance, for the phrase AADAA UMMATAA OROMOO is generated in a way that AADAA AA D AH, UMMATA UH M AH T AH and OROMOO AO R AH M UW. Some other need manual edition of the phonemes generated such as QONNAA K AA N AH and CAAMSAA K AA M Z AH needs generated are not correct as the phonemes for Afaan Oromoo different from English. Therefore the two above generated lists of phones are edited manually in such a way that QONNAA Q OO N AH and CAAMSAA C AA M Z AH respectively. Furthermore the generated phones are Q, OO AH, AA, M, Z from the two words selected from the phrases given in the corpus.

5.2.2. Feature vector extraction

Acoustic models are created by taking audio recordings of speech and their text transcriptions. These components are then processed to create statistical

representations of the sounds that make up each word. The acoustic model is then used by the speech recognition decoding engine to actually recognize speech.

As a result of this, acoustic models are essential components to a speech recognition system. Here we can say even the performance of the recognizer is as good as its acoustic model.

Hence creating the feature vectors is most important aspect for creating the acoustic and language model. With this regard one of the predominant feature vector file format is, the MFCC feature files, generated by sphinxtrain for the training data and the entire models was created in accordance with the preceding steps of the model development. The MFCC file format the configuration file sphinx_train.cfg with its default parameters was created and used with slight modifications. In particular the feature files were created by the script given below:

Perl ../scripts_pl/make_feats.pl –ctl etc/afaanoromoo_train.fileids

The features files created using the above script are then used for latter decoding with parameter adjustment and the model trained whenever there is no error on the training data. Here whenever there are problems at this step the log file will assist us in finding and solving those problems. The feature files generated are used for training which is elaborated in the training of the HMM section.

5.2.3. Dictionary preparation

Dictionary preparation is yet another important task to be performed for the model. For the model under construction the prepared dictionary should be in line with the

transcription file and file identifications (fileids) of the wav files and uses all sorts of activities performed in the etc folder as per the procedure.

For the two models, the word level model and phoneme level model, the dictionaries were prepared separately. For word level the dictionary takes the entire phrases and words as it is. Whereas, for Triphone based model the dictionary was performed which maps the words with their respective phones. An sample of the dictionary constructed and used for triphone level phoneme based are shown in the following table 5.3

Phoneme based model of phrases	
DHUGAATII	D HH Y UW G AH T IY
DIINAGDEE	D IY N AH G D E E
DUBARTOOTAA	D Y UW B AH R T UW T AH
GARAAGARRUMMAA	G AE R AH G AH R AH M AH

Table 5.3. Sample dictionaries constructed and used for the model

In addition to the dictionaries of the words, phrases and sentences in the corpus there are also filler dictionaries used to separate and indicated the point at which the speech started, pause and end. This is mostly indicated using the silent at beginning, silence in between utterances, and silence at the end of the utterance. The filler dictionaries used are indicated in table 5.4.

Filler dictionary content	**Purposed used for in the dictionary**
<s> SIL	Silence at the beginning of words/phrases and sentences
<sil> SIL	Silence in between words/phrases and sentences
</s> SIL	Silence at the end of words/phrases and sentences

Table 5.4. Filler dictionary used for CAOSR

5.2.4. Training the HMM

Training the model is performed with sphinx train tool available for sphinx system and some task are performed before the actual training is done. Some pre-processing tasks were performed before the actual training conducted including, cygwin bash shell installation so as to use the command environment of the windows (also possible to use linux, unix terminals). The sphinxtrain and our model (Afaanoromoo) were created in the same place so as to facilitate the model for training.

After creating the two directories, in the same place, we need to create the task in the task directory. The task directory is created by navigating the path to the task directory (Afaanoromoo) and running the script as follows:

perl ../sphinxtrain/scripts_pl/setup_sphinxtrain.pl –task Afaanoromoo

The above command executed that produced the subdirectories and the necessary configuration files used for latter steps. For instance, the wav directory is for putting raw audio files where as the etc directory is created having important files such as configuration file sphinx_train.cfg which is used by editing the content. For this reason as the model created the default path and the default file extensions of the Sphinxbase, we need to provide the appropriate paths and file extensions that we had. In accordance with the sphinx training for Afaan Oromoo language was created by the following scripts by directing the directory to the sphinx train.

Perl ../scripts_pl/Runall.pl

After the scripts was run which takes up to 30 minutes so as to produce the acoustic model for Afaan Oromoo language. Hence, this is the step at which the actual acoustic model of the language was created. The log file is the file that is automatically created in the task directory with html format so as report the tasks performed.

5.2.5. Language model

The language model predicts the most likely continuation of an utterance on the basis of statistical information about the frequency in which words, phrases and the sentences sequences occur on average in the language to be recognized. Like the HMMs, an efficient language model needs be trained on large amounts of data, in this case texts collected from the target is used.

Accordingly, the language model for Afaan Oromoo languages used in this experiment is generated using the Sphinx Knowledge Base. The language model prepared for the recognizer is then taken to the JAR files to integrate and the language model for all the three 1 gram, 2 gram and 3 gram were prepared for the recognizer and used accordingly.

Sample language model for a given phrase can be given the following manner as representing the initial probability at the beginning and back-off weight given in log 10 at the end of the phrase.

```
-0.3010 CAASAA MOOTUMMAA -0.1249
```

Accordingly, the two language models were developed for the monophone word level and triphone phoneme level. The word level language model takes all the words and or phrases that are found in the corpus.

5.2. Preparation for Decoding

The decoder is, simply to put the decoder algorithm that tries to find the utterance that maximizes the probability that a given sequence of speech sounds corresponds to that utterance. As per the literatures, it is a search problem, and especially in large vocabulary systems careful consideration must be given to questions of efficiency and optimization, for example to whether the decoder should pursue only the most likely hypothesis or a number of them in parallel (Young, 1996).

According to (Young,1996), an exhaustive search of all possible completions of an utterance might ultimately be more accurate but of questionable value if one has to wait two days to get a result. Trade-offs are therefore necessary to maximize the search results while at the same time minimizing the of CPU and recognition time.

The model trained using sphinxtrain is prepared for decoding and testing with sphinx4 decoder. Up on adjusting the parameters of the acoustic and language models of the Afaan Oromoo Continuous Speech Recognizer is transferred to the appropriate directories and performed the necessary steps for decoding.

Preparing of the training model for decoding is performed by integrating the components of the sphinx4 and the JAR files which the java components are created by ant scripts in the performance directory. Here the accuracy tracker is prepared for sphinx system is used by making necessary parameter setting and components of the recognizer decoding model.

Finally the accuracy tracker performs the accuracy level obtained by comparing the batch file which is the texts to be recognized with the with the audio test sets in raw format. To achieve this, the ant scripts executed and a performance was

automatically generated with its respective statistical summary. In the next part, the result for the recognizer with their discussion is presented. The components of the JAR file used to integrate the components of the recognizer are included in annex C:

5.3. Analysis and Discussion of the Experiment Result

The performance of the speech recognition system is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of performance accuracy which is usually rated as Word Error Rate (WER), whereas speed is measured with the real time factor (Jurafsky et. al., 2000).

Results output by any ASR system have to be compared with expected correct results in order to measure system's performance. To this end one of the important measure of performance is the word correct rate and computed by the formula as;

$WordCorrectRate = 100\frac{N - D - S - I}{N}$ Where N denotes the total number of words in recognized sentences, D denotes deletions, S denotes substitutions and I refer to Insertion errors.

Substitution errors occurs when an utterance in the target vocabulary is misrecognized as another vocabulary item, on the other hand deletion errors occur when a target vocabulary item is not deleted where as insertion errors occurs when non target is recognized as the target item.

According to the experiment conducted for Afaan Oromoo Continuous, speaker independent Speech Recognizer, which is conducted for both context dependent word level and context independent, the performance levels were obtained.

The task performed for both context independent and context dependent which are described independently in the following section. For that matter, the two models have different phones, language models, pronunciation dictionary, and other components so as to compute the respective performance level. As a result, the experiment conducted and the results obtained for both models. The following table 5.5 shows the result for context dependent model.

Summary Statistics of the Recognizer	
Yardsticks	**Values**
Words	700utterances(1480 words)
Errors	963(Sub: 430,Ins:497, Del:36)
Word accuracy	68.514%
Sentence accuracy	28%
Time	Audio 963.58s proc:602.12s
Speed	0.62 X real time
Memory	This :115.16Mb Avg.144.76Mb Max.206.18Mb

Table 5.5 recognizers performance result for the context dependent model

According to table 5.5., the experiment result for word level continuous Afaan Oromoo speech recognizer indicates 700 test data sets and of which contain 1480 different word and showed a performance level of 68.514% with sentence accuracy of 28% which is promising for continuous utterances. The errors indicated are more or less covered by the substitution and insertion errors. This is due to the nature of the

environment used and the co articulation effects. Some word and sentence are related and the substitution and insertion errors accounting nearly 96% of the errors committed by the recognizer model. Yet another experiment is conducted which resulted in better performance than the word level.

For context independent, another experiment performed and the performance level was obtained. The experimental result summary for context independent is given the following table 5.5.

Summary Statistics of the Recognizer	
Yardsticks	**Values**
Words	700(utterances) (1480words)
Errors	176(Ins: 116,Del:20, Sub:40)
Word accuracy	89.459%
Sentence accuracy	42%
Time	Audio:963.58s processing:95.35s
Speed	0.10 X real time
Memory	This:39.51MbAverage:68.55MbMax:122.79Mb

Table 5.6. Recognizer performance for context independent model

According to the above table the performance for context independent is found to be 89.459% and the sentence accuracy was 42% and this indicates the result obtained from the context independent which is a phoneme based was found to be good result.

The errors indicated are 176(Insertion: 116, Deletion: 20, Substitution: 40) and of which the largest share goes to the substitution errors of the total nearly 66% of the

errors observed. For example from the accuracy tracker it is found to be the word SIRBA AADAA is substituted by AADAA for one speaker, while BALA KONKOLAATA is substituted by BALA KITAABAA for another.

According to the experiments conducted, the sentence accuracy is 42% and 28% respectively for context independent and context dependent models. This in turn indicates the context independent phoneme level by far better than the context dependent one. Not only the accuracy level, but it is also possible to not the see the speed, time and memory usage of the two models.

The performance of the two models were checked against the level of accuracy obtained and the efficiency in execution time and other factors of the prototype, and the context independent phoneme level is found to be the best for Afaan Oromoo continuous, speaker independent speech recognizer.

Furthermore, when we compare the two models, the context independent model by far out weights the context dependent model. Here the context independent model which is a phoneme level out performed and the result is promising. Lastly, in the next chapter the conclusions and recommendations for the recognizer performed are presented.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

Since continuous Afaan Oromoo speech recognizer prototype development is performed and some improvements have been achieved. The following chapter intends to show the concluding remarks have been made and further future works are forwarded.

6.1. Conclusions

As the intension of every speech recognizer is to develop the model that converts speech utterances to texts, continuous, speaker independent Afaan Oromoo speech recognizer is also performed towards achieving the objectives sited earlier and in general to convert continuous Afaan Oromoo speeches to texts.

This thesis work is performed through integration and implementation of appropriate tools, techniques and methodologies. Furthermore, the special features of the language are also entertained due to the speech characteristics taken in to considerations, so as to arrive in to the appropriate results for the language under investigation.

Mainly the purpose of this research was towards exploring the possibility and applicability of the selected tools and techniques to continuous Afaan Oromoo speech utterances and the possible outcomes that recognizer brought up.

To arrive to this, the researcher performed different activities towards achieving the objective. Accordingly the open source tool from Sphinx System, Sphinx Train and Sphinx4, was selected which work with the statistical approach under HMM for continuous Afaan Oromoo speech recognizer as the tool have a variety of modules that implement state-of-the-art speech recognition techniques beyond its modularity and flexibility.

As to the knowledge of the researcher, this research is the first of its kind for continuous Afaan Oromoo speech recognition where as one attempt for isolated word recognition for Afaan Oromoo made earlier.

Furthermore, the researcher prepared two language models for word level context dependent and triphone which is a context independent so as to check the difference in the recognizer model. Accordingly, research output for medium vocabulary continuous Afaan Oromoo speech recognizer for word level and triphone based tried and the performance level of 68.514% with sentence accuracy of 28% and 89.459% with sentence accuracy of 42% were achieved for both respectively.

It is also important to note that the performance of triphone based context independent out performs by far the context dependent monophone recognizer models. This is mainly because as the model developed is a continuous speech recognizer for Afaan Oromoo, the context dependents performance became less due to the factor that the recognizer performs on the phrases and sentence and it is as far as each and every phrase is found that the performance of the recognizer becomes

higher. Therefore it is clear that the performance of recognizer is become less compared to context independent.

In addition to this as this research is performed for speaker independent continuous speech recognizer for Afaan Oromoo, the impact of speaker independence is seen significantly. Hence the work for speaker dependent hopefully will improve the performance of the work.

The performance of continuous speech performance is promising in relation to the other related researches in the area in other languages including Amharic and Tigrigna. As this is the first research for continuous Afaan Oromoo speech recognizer in the future the consideration of other factors might increase the performance beyond this.

According to the result obtained from the model constructed it is found to be good performance level in comparison to those related works seen so far. In addition to this, the researcher decided to put some recommendations for further research work to be conducted latter in the area.

6.2. Recommendations

Up on the result obtained from the developed prototype, the researchers finding is found to be promising while some others need to be recommended for further researches in the future so as to enhance this work or to further extend the full fledged recognizer for Afaan Oromoo.

Apart from the trial for developing whether the isolated word or continuous speech for Afaan Oromoo, the language as it has no corpus prepare for this purpose, preparing the corpus itself can be one of the research areas in the domain. Therefore we recommend further research to this aspect of the language.

Hence the need for investigating the speaker adaption is one of the disciplines which need further investigation in the language Afaan Oromoo, the need for conducting research in this area also is one of the important aspects to be considered.

The need to consider the different dialects to arrive at appropriate recognizer in the language is one of the important factors to be considered as the language Afaan Oromoo is rich in dialectic variation.

The environment in which the speech corpus is recorded was controlled. Even though the environment is controlled by the researcher to make silent environment, it still needs further silent environment in which the recording to be conducted. The speakers also show variations in uttering the words, phrases and sentences which resulted in lack of uniformity exerting significant effect on the recognizer. To

minimize such variations, the speakers have to be trained how to utter the speech data set.

As the size of the speech corpus increases, the performance of the recognizer also increases the accuracy level. Hence, a large corpus should be considered in the future works. As a result of the increase in corpus the need for other requirements in the capability of the machine to run those data is also important.

Lastly as this research is conducted on read speech, the need to consider spontaneous speech, speeches from TV broadcast, Radio and conference halls might be considered in the future works.

Reference:

A. Mohammad, 2006, isolated word automatic recognition system in the telephony, MSc thesis, university of Malaya, Kuala Lumpur, Malaysia.

Asafa Jalata, 2010, the present and future of the Oromoo people, Journal of the Oromoo literature, a scholarly publishing initiatives, Minnesota, USA

Ashenafi Demissie, 2009, A Speech Recognition System for Afaan Oromoo, MSc thesis, Addis Ababa University, Addis Ababa.

Bahi, L. R. ,J.K. Baker, P.S. Cohen, A.G. Cole, F. Jelinek, B.L. Lewis, and R.L. Mercer,(1978), Automatic Recognition of Continuously Spoken Sentences from a Finite State Grammar, In Proc ICASSP, pp. 418-421.

Bernd Planner, 2005, Introduction to Speech Recognition –Edition1.1, Munich-German, IEEE publications.

Daniel Jurafsky and James H. Martin, 2000 .An introduction to Natural Language processing, computational Linguistics , and speech Recognition, University of Colorado prentice Hall Inc international edition pp. 235-283.

Edward C. Lin, Kia Yu, Rob A. Rutenbar, Tsuhan Chen, 2006 A 1000-word vocabulary, speaker-Independent , Continuous Live-Mode Speech Recognizer implemented in a single FPGA, article Carnegie Mellon University Pittsburgh U.S.A.

Furui, S. (1986) Speaker Independent Isolated Word Recognition using Dynamic Features of the Speech Spectrum. *IEEE Trans on Acoustics, Speech and Signal Processing*, Vol. 34, No. 1, pp.52-59.

Ganapathiraju, A. (2002). Support Vector Machine for Speech Recognition, Ph.D Thesis, Carnegie Mellon University.

Gray, R. *Vector Quantization.* IEEE ASSP Magazine, vol. 1 (1984), pp. 4-29.

Hope L.Doe, 1998, Evaluating the Effects of Automatic Speech Recognition Word Accuracy, MSc Thesis, Blacksburg, Verginia, USA.

Hualin gao, richard duncan, julie a. baca, joseph picone, 2002, signal processing tools for speech recognition, journal of information science, Mississippi state University

Huang, X. and Jack, M. *Semi-Continuous Hidden Markov Models with Maximum Likelihood Vector Quantization,* IEEE Workshop on Speech Recognition. 1988.

International Phonetic Association, *Handbook*, pp. 194–196 available at: http://www.omniglot.com/writing/ipa.htm (visited at: Monday march 15, 2010)

J. Sakai and S. Doshita, The Phonetic Typewriter, Information Processing 1962, Proc. IFIP Congress, Munich, 1962.

J. Suzuki and K. Nakata, Recognition of Japanese Vowels—Preliminary to the Recognition of Speech, J. Radio Res. Lab, Vol. 37, No. 8, pp. 193-212, 1961.

J. W. Forgie and C. D. Forgie, Results Obtained from a Vowel Recognition Computer Program, J. Acoust. Soc. Am., Vol. 31, No. 11, pp. 1480-1489, 1959.

Juang and Rabiner, 2004, Automatic speech Recognition – A brief history of the technology Development,Georgia Institute of Technology, Atlanta, Rutgers,University and the University of California, Santa Barbara, article.

Juang, B. H. & Rabiner, L. R. (1991). Hidden Markov Models for Speech Recognition. Technometrics, vol. 33, no 3, pp. 251-272, 1991.

Juraj Kačur ,2004,white paper "HTK vs. SPHINX for SPEECH Recognition" by Ilkovičová 3, Bratislava, Slovakia(visited at: 19/02/2010)

K. H. Davis, R. Biddulph, and S. Balashek, Automatic Recognition of Spoken Digits, J. Acoust. Soc. Am., Vol 24, No. 6, pp. 627-642, 1952.

K.H.Davis, R.Biddulph, and S.Balashek, Automatic Recognition of spoken Digits, J. Acoust. Soc. Am.,Vol 24,No. 6, pp.627-642,1952.

Kimberele,A.K.,2003" An Introduction to speech Recognition " IBM Corporation. URL:http://www.ibm.com/software/pervasive/products/pdf/introduction_to_speech_Recognition.pdf [visited at: 18/12/2009].

Kurzweil, 2002, Developing continuous speech recognition technology that uses natural language processing commands.

Lewrence R. Rabiner, 1989, a tutorial on Hidden Markov Model and selected applications in speech recognition, preceeding of IEEE ,vol.77 Number 2, 1989

M. Honda, NTT CS Laboratories, Speech synthesis technology based on speech production mechanism, How to observe and mimic speech production by human, Journal of the Acoustical Society of Japan, Vol. 55, No. 11, pp. 777-782, 1999

M. Strauss F. Jelinek, B. Merialdo, and S. Roukos, 1992 A Dynamic Language Model for Speech Recognition, IBM research division.

Martha Yifru, 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.

Mei-Yuh Hwang, 1993 subphonetic acoustic model for speaker independent continuous speech recognition, PhD dissertation, Carnigie Mellon University,USA

Michael K. Brown, 2001, Stochastic Language Models (N-Gram) Specification, Avaya Labs, article paper.

Nebiyou Tsegaye, 2005, speech to text conversion using Amharic characters, MSc thesis Addis Ababa University, Faculty of technology, Department of communication engineering, Addis Ababa University.

S. E. Levinson, L. R. Rabiner, and M. M. Sondhi,(1983) An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition, Bell Syst. Tech. J., Vol. 62, No. 4, pp. 1035-1074, April 1983.

Sami Lemmetty, 1999, Review of speech synthesis technology, a Master's thesis, Helsinki University of technology, Finland.

Stephen Cook, 2002, Speech recognition –How to. Available: http://www.faqs.org/docs/Linux-HOWTO/Speech-Recognition-HOWTO.html (visited at: 18/12/2009).

Taha Roba and C Wilson-Owens, (2003) Alphabets and sounds – sagaleewani fi loqoda. Available at http://www.Ethnomed.org (visited at: Monday ,march 15,2010)

Tan chin Luh, 2004, speaker independent speech recognition using neural Network, MSc thesis Universiti Putra, Malaysia

Tebelskis, J. (1995). S*peech Recognition using Neural Networks*, Ph.D. Dissertation, Carnegie Mellon University.

Tilahun Gamta, 1992,The Oromo language and the latin alphabet, Journal of Oromo studies – Addis Ababa University.

Ting Chee Ming(2007) Malay continuous speech recognition using continuous density hidden markov model, MSc thesis, faculty of electrical engineering universiti teknologi malaysia,

Wiggers Paskal, 2001, Hidden Markov Model for Automatic Speech Recognition and their multimodal Applications, Delft University of technology, the Netherlands.

Zegaye Seifu, 2003,Hidden Morkov Model Based Large Vocabulary, Speaker Independent, continuous Amharic Speech Recognition, MSc thesis Faculty of Informatics, Addis Ababa University, Addis Ababa.

Appendix

Annex A: Sample trigram Language Model used for the Experiment

```
\data\

ngram 1=127

ngram 2=202

ngram 3=148

\3-grams:

-0.3010 <s> AADAA UMMATA

-0.3010 <s> ABDII BORUU

-0.3010 <s> ADDA BILISSUMMAA

-0.3010 <s> ADDA-DUREE </s>

-0.3010 <s> AKKA LAKKOOFSA

-0.3010 <s> AKKAATAA JIREENYAA

-0.6021 <s> AYYAANA GARII

-0.6021 <s> AYYAANA QILLEE

-0.3010 <s> BAGA ITTIIN

-0.3010 <s> BALAA KONKOLAATAA

-0.3010 <s> BARMAATILEE MIIDHAA

-0.3010 <s> BIQILTOOTA DHAABUU

-0.3010 <s> CAASAA MOOTUMMAA

-0.3010 <s> CHAAPPAA BARBAADI

-0.3010 <s> DAANDII BAADIYYAA

-0.3010 <s> DHAAMSA DABARSUU

-0.3010 <s> DHUGAATII AADAA

-0.3010 <s> DIINAGDEE FI

-0.3010 <s> DU'AA KA'UU

0.3010 <s> ULAAGALEE BARBAACHISAN

-0.3010 <s> ULAGAA GUUTUU

-0.3010 <s> UMMATA BAADIYYAA

-0.3010 <s> UMMATTOOTA KIBBAA

-0.3010 <s> WAAJJIRA MOOTUMMAA

-0.3010 <s> WAL-QUNNAMTII SAALAA

-0.3010 <s> WALDAA DUBARTOOTAA

-0.3010 <s> WALDAALEE XIXIQQAA

-0.3010 <s> WEELLISTOOTAA OROMOO

-0.3010 <s> XALAYAA HOJII

-0.3010 <s> YAKKARRAA BILISA

-0.3010 AADAA UMMATA OROMOO

-0.3010 ABDII BORUU </s>

-0.3010 ADDA BILISSUMMAA </s>

-0.3010 AKKA LAKKOOFSA ITOOPHIYAA

-0.3010 AKKAATAA JIREENYAA </s>

-0.3010 AYYAANA GARII </s>

-0.3010 AYYAANA QILLEE </s>

-0.3010 HAWWII GAARII </s>

-0.3010 HUBANNOO BALDHISUU </s>

-0.3010 ISIN GAHE </s>

-0.3010 ITTIIN ISIN GAHE
```

-0.3010 <s> FUDURAA FI	-0.3010 JAALALA DHUGAA </s>
-0.3010 <s> FUUDHA FI	-0.3010 JAL’ISII AMMAYYAA </s>
-0.3010 <s> GARAAGARRUMMAA </s>	-0.3010 JEEQUMSA BARATTOTAA </s>
-0.3010 <s> GEEJJIBA AMMAYYAA	-0.3010 JIBBITUS JAALATTUS </s>
-0.3010 <s> GUMII SHAMARRANII	-0.3010 JIRUUFI JIREENYA </s>
-0.3010 <s> GURMAA’INA DARGAGGOOTAA	
-0.3010 <s> HARA DAMBAL	-0.3010 WAL-QUNNAMTII SAALAA </s>
-0.3010 <s> HAWWII GAARII	-0.3010 WALDAA DUBARTOOTAA </s>
-0.3010 <s> HUBANNOO BALDHISUU	-0.3010 WALDAALEE XIXIQQAA </s>
-0.3010 <s> JAALALA DHUGAA	-0.3010 WEELLISTOOTAA OROMOO </s>
-0.3010 <s> JAL’ISII AMMAYYAA	-0.3010 XALAYAA HOJII </s>
	-0.3010 YAKKARRAA BILISA TA’UU
	\end\

Annex B. Selected Afaan Oromoo phrases and sentence used for Experiment

AADAA UMMATA OROMOO	GARAAGARRUMMAA
MOOTUMMAA NAANNOO OROMIYAA	CAASAA MOOTUMMAA
DIINAGDEE FI HAWAASA	JI’A CAAMSAA
JIRUUFI JIREENYA	DU’AA KA’UU
MALAAMMALTUMMAA BALLEESSUU	JEEQUMSA BARATTOTAA
OROMIYAAN HAADAGAAGDUU	JAL’ISII AMMAYYAA
AKKA LAKKOOFSA ITOOPHIYAA	FUDURAA FI KUDURAA
MATA-DUREE ODUUWWANII	DHAAMSA DABARSUU
BIQILTOOTA DHAABUU	MANA KITAABAA
GURMAA’INA DARGAGGOOTAA	LAGA GAMA
MANNEEN MOOTUMMAAN IJAARE	GUMII SHAMARRANII
MANA BARUUMSA QOPHAA’INAA	LUBBU QABEEYYII
LAKKOOFSA UMMATA ITOOPHIYAA	QOPHII ADDAA
MOOTIIWWAN ITOOPHIYAA	XALAYAA HOJII
MANA NYAATAA	HUBANNOO BALDHISUU
BARMAATILEE MIIDHAA QABAN	NYAATA AADAA
DAANDII BAADIYYAA	ABDII BORUU
KAROORA MAATII	ULAAGALEE BARBAACHISAN
JAALALA DHUGAA	ULAGAA GUUTUU
HAWWII GAARII	WAAJJIRA MOOTUMMAA
WAL-QUNNAMTII SAALAA	MALA JIREENYAA
QOTE-BULAA	SAB-QUNNAMTII
LAFA QONNAA	AYYAANA QILLEE
FUUDHA FI HEERUMA	GEEJJIBA AMMAYYAA
SHAMARRAN OROMOO	DHUGAATII AADAA
AYYAANA GARII	YAKKARRAA BILISA TA’UU

MANA BARUMSAA	SIRBA AADAA
CHAAPPAA BARBAADI	WEELLISTOOTAA OROMOO
ADDA BILISSUMMAA	AKKAATAA JIREENYAA
BAGA ITTIIN ISIN GAHE	WALDAALEE XIXIQQAA
BALAA KONKOLAATAA	MATA-DUREE
UMMATA BAADIYYAA	ADDA-DUREE
LUKKUU QALUU	JIBBITUS JAALATTUS
UMMATTOOTA KIBBAA	
HARA DAMBAL	
LAGGEEEN OROMIYAA	
WALDAA DUBARTOOTAA	

Annex C: Sample scripts to integrate the components of JAR files

```
    <property name="afaanoromoo"
value="WSJ_8gau_13dCep_16k_40mel_130Hz_6800Hz"/>

    <property name="wsj_data_dir" value="models/acoustic/wsj"/>

    <property name="wsj_8kHz_name"
value="WSJ_8gau_13dCep_8kHz_31mel_200Hz_3500Hz"/>

    <property name="wsj_8kHz_data_dir" value="models/acoustic/wsj_8kHz"/>

    <property name="tidigits_name"

              value="TIDIGITS_8gau_13dCep_16k_40mel_130Hz_6800Hz"/>

    <property name="tidigits_data_dir" value="models/acoustic/tidigits"/>

    <property name="afaanoromoo_name"
value="AFAANOROMOO_8gau_13dCep_16k_40mel_130Hz_6800Hz"/>

    <property name="rm1_data_dir" value="models/acoustic/rm1"/>

<property name="afaanoromoo_name"
value="AFAANOROMOO_8gau_13dCep_16k_40mel_130Hz_6800Hz"/>

<property name="afaanoromoo_data_dir" value="models/acoustic/afaanoromoo"/>
```

DECLARATION

I, THE UNDERSIGNED, DECLARE THAT THIS THESIS IS MY ORIGINAL, HAS NOT BEEN PRESENTED FOR A DEGREE IN ANY OTHER UNIVERSITY AND THAT ALL SOURCES OF MATERIAL USED FOR THE THESIS HAVE BEEN FULLY ACKNOWLEDGED.

NAME SIGNATURE DATE

CONFIRMED BY ADVISOR:

NAME: _____SIGNATURE: _____

DATE: _____