ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULITY OF SOCIAL SCIENCE AND HUMANITIES

DEPARTMENT OF LINGUISTICS

Handling Pronunciation Variation Using Hybrid Approach in Continuous, Speaker Independent Speech Recognition for Amharic

By

SHIMEKIT TEKA

OCTOBER, 2014

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULITY OF SOCIAL SCIENCE AND HUMANITIES

Handling Pronunciation Variation Using Hybrid Approach in Continuous, Speaker Independent Speech Recognition for Amharic

ADVISERS: SOLOMON TEFERA (PHD)

FEDA NEGESSE (PHD)

BY

SHIMEKIT TEKA

OCTOBER,2014

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULITY OF SOCIAL SCIENCE AND HUMANITIES

Handling Pronunciation Variation Using Hybrid Approach in Continuous, Speaker Independent Speech Recognition for Amharic

BY

SHIMEKIT TEKA

Signature of the Board of Examiners for Approval

Name	Title	Signature	Date
_____	Chair Person	_____	_____
Solomon Tefera	Advisor	_____	December 23, 2014
Teda D.	Advisor	_____	December 23, 2014
_____	Examiner	_____	Dec. 23, 2014
Dereb A	Examiner	_____	24 Dec 2014

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULITY OF SOCIAL SCIENCE AND HUMANITIES

Handling Pronunciation Variation Using Hybrid Approach in Continuous, Speaker Independent Speech Recognition for Amharic

BY

SHIMEKIT TEKA

Signature of the Board of Examiners for Approval

Name	Title	Signature	Date
_____	Chair Person	_____	_____
_____	Advisor	_____	_____
_____	Advisor	_____	_____
_____	Examiner	_____	_____
_____	Examiner	_____	_____

DEDICATION

This thesis is dedicated to my wife Tewabech Beyene and our son Meklit Shimekit.

Acknowledgements

Glory is to the Almighty God that I came to this point in life. Next, I wish to express my deepest gratitude to my advisors, Dr. Solomon Teffera and Dr. Feda Negesse.Their guidance, inspiration and advice with untiring and selfless help throughout my study finally led me to successfully complete my thesis.

Next, I am deeply indebted to my friend Onosmos Amberas who was with me in all aspects from the early beginning of this research to the end. Fetsum Seyoum, my friend and colleague, gave me ideas whenever we met for tea, on the phone or just when we passed by each other anywhere in the campus.

My beloved wife, Tewabech took the burden at home right after a year of our weddingwhen I joined this department. She endured the loneliness of some unforgettable days and nights when I have been here. I would like to show my sincere gratitude to my parents, brothers and sisters who have always encouraged me to further my studies and become a better person.

I am grateful to East Badawacho woreda Education Office for the financial assistance that cover this program.

Above all, I thank God for everything that He has done for me, without which I couldn't have done anything.

TABLE OF CONTENTS

List of Abbreviations

ASR: ..... Automatic Speech Recognition

HMM: ..... Hidden Markov Mode

HTK: ..... Hidden Markov Model Toolkit

LPC: ..... Linear Predictive Coding

MFCC: ..... Mel Frequency Cepstral Coefficient

LPCC: ..... Linear Predictive Cepstral coefficient

PLP: ..... Perceptual Linear Prediction

OCR: ..... Optical Character Recognition

LG: ..... Language Model

I..... insertion

D.....deletion

S.....substitution

H.....hits

List of tables and figures

Abstract

The problem of modeling pronunciation variation lies in accurately predicting the word pronunciations that occur in the test material .In order to achieve this, the pronunciation variants must first be obtained in some way or other i.e. from pronunciation data or from pre-specified phonological rules based on linguistic knowledge.

In this study, first models was developed using canonical dictionary as a reference with total data set of 950 sentences of which 700 for training and the remaining 250 are used for testing model

Two models were developed using knowledge based and data driven adding variants to the respective dictionaries. Lastly another model: hybrid approach was developed. This model is supposed to avoid the short coming of the two models, knowledge based and data driven. In light of this, the model developed using hybrid approach has shown better performance. The better progress obtained is due to the fact that in case of knowledge based approach, the variant added may or may not really appear in the text audio. in contrary, during data driven the variants actually appear in audio but it is very difficult to listen and identify all words with variation. Thus the experiment undertaken in this study revealed that combining the two approaches (hybrid) is the best way for handling pronunciation variation.

CHAPTER ONE

INTRODUCTION

1.1. Background

The field of speech signal analysis has been in the center of attention for many years because of the many possible applications which enable human beings to exchange information using the most natural and efficient way. Besides, the use of such recognition System requires no training. Furthermore, since we speak much faster than we write and type, speech provides the highest potential capacity in human-to-machine communication. In this context the speech recognition enterprise is probably the most ambitious. The goal of speech recognition is to build intelligent machines that can hear and understand spoken information, in spite of the ambiguity and complexity of natural languages.

Spoken Language is the fundamental part of everyday communication among humans for a variety of reasons in our life (Kessen, 2002). As a result, many people think that speech might also be a very efficient means of communication between human beings and machines. The tremendous growth of technology increased the need of integration of spoken language technologies into our daily applications, providing an easy and natural access to information.

Humans have been dreaming of and struggling for developing intelligent machines which master speech input and output to communicate with them naturally. In other words, the machines have to understand human speech and generate human like speech (Chollet, 1994). As pronunciation relates to speech, humans have an intuitive feel for pronunciation.Words are pronounced differently by different people, a phenomenon called "pronunciation variation." Pronunciation variation has been studied in thespeech recognition field (Crème,1996). Pronunciation variation is a phenomenon observed within a speaker or within a group of speakers of the same dialect or

among speakers across dialects of the same language. Pronunciation variation occurs as there are different ways of saying a given word (cited by Solomon, 2005). Modeling of pronunciation variability is an important part of the acoustic model of a speech recognition system to handle variation that might decrease the ASR performance. Good pronunciation models contribute to the robustness and portability of a speech recognizer *(Hain, 2000).*

Attempts have been made in the field of speech synthesis and recognition to handle pronunciation variation and improve performance of the corresponding speech systems. A first major distinction in pronunciation variation can be drawn between intra-speaker and inter-speaker variations. Intra-speaker variation refers to the fact that the same speaker can pronounce the same word in different ways depending on various factors. The first important factor that may affect the way in which words are pronounced is the fact that "they are strung together into connected speech" (Kaisse, 1985) as opposed to when they are pronounced in isolation. In connected speech, all sorts of interactions may take place between words, which will result in phonological processes such as assimilation reduction, deletion and insertion. The degree to which these process occur will vary depending on the style of speaking (Amdal and Eric fosler - Lussier 1989).Early automatic speech recognition (ASR) systems only considered restricted speaking styles, i.e. acareful articulation of isolated or connected words. The increased modeling capacities of current ASR systems also manage the articulation of continuous speech.

1.2. Statement of the Problem

The ultimate goal of any research in speech recognition is to develop a system that can understand the speech of anyone who needs to use the system.To meet this requirement, the system must be able to handle inter-speaker and intra-speaker variability, recognize unrestricted

vocabulary and understand fluent, conversational speech Wester(1971). Despite the fact that highly developing sophisticated technology, this could not fully achieved due to certain constraints. One of these constraints which remained bottle neck to the successful performance of the ASR system is pronunciation variation.

Pronunciation is the way a word or a language is customarily spoken. Humans have intuitive feeling for pronunciation (Lee, 2006).When a word is pronounced in different ways, humans can tolerate the variation. By contrary, machines do not tolerate when one word is pronounced differently by the same individual or group of individuals.This variation in pronunciation is a major problem in ASR as it impedes ASR recognition system.

In Amharic we find words being pronounced in different ways which is variation due to intra-speaker (within the same individual) or inter-speaker (between individuals) which may be one of the factors for Amharic ASR poor performance. Solomon (2013) and Tewedaj (2013) have conducted research in handling pronunciation variation using knowledge based and data-driven approach respectively. The problem of knowledge-base is that ,the information from linguistics literature is not exhaustive ;many process that occur in real speech are yet to be described . Besides the rules are often very general and thus many variants are generated or not enough variants might be generated .Further more no information on how often the generated variants appear in the data (Young,2002). On the other hand the problem of data-driven is that it is extremely difficult to extract reliable information from the data; that is ,the information which is used to generate variants is not always reliable. Besides it is very much data base dependent and variants that occur frequently in one speech corpus do not necessarily occur frequently in other corpora (Young,2002).

Both have recommended hybrid approach to investigate the relative performance of this approach. Therefore, this research aims at handling pronunciation variation due to inter-speaker by using hybrid approach.

1.3. Research Questions

This study answers the following questions:

1. Does hybrid approach relatively show better improvement in handling pronunciation variation?
2. How can we handle pronunciation variation in Amharic ASR system?
3. How does pronunciation variation affect ASR performance?

1.4. Objective of the Study

1.4.1. General Objective

The general objective of the study is to explore way of handling pronunciations variation by using both knowledge-based and data-driven in continuous, speaker independent speech recognizer for Amharic.

1.4.2. Specific Objective

In order to achieve the general objective, the following specific objectives are set.

- To investigate causes of pronunciation variation.
- To prepare data set for training and testing.
- To capture and exploit how pronunciation variation affects ASR performance.
- To develop alternative pronunciation variation
- To develop ASR system using different pron. Dictionaries

- ❖ Evaluate the performance of the ASR systems

1.5. Significance of the Study

The development of natural language processing especially automatic speech recognition has wide application with the highly developing technology. But due to different constraints which hinder good performance of the ASR system, using this technology for the intended objective is quite difficult especially for local languages. One constraint for ASR poor performance is pronunciation variation.Therefore, the significance of this study is to identify thesource of pronunciation variation attributed for ASR poor performance and paving way how to handle this problem in order to improve ASR performance for Amharic. Campanies increasingly use and develop system with speech recognition interfaces citing various benefit including(Marcowitiz 1996):

- Increase productivity by enabling a person to use her/ his hands and mouth for different tasks and making hands free work possible.
- Rapid return on investments that apply ASRSs to speed up tasks,
- Environment control(by disable people),
- The naturalness of communication between man and machine.

1.6. Scope and Limitation of the Study

The scope of this research was confined to investigating how to handle pronunciation variation in speaker independent, continuous speech recognizer for Amharic. The study did not consider different accent across regional dialect of Amharic; it focused on, the standard dialect spoken in Addis Ababa. Different constraints limited the effectiveness of this study. Among these listening and identifying words with variants during data driven approach was one major constraint. The

other constraint is some of the variants added to the dictionary during knowledge based approach do not appear in actual sense and this may create confusion to the search algorithm.

1.7. Methodology

Methodology refers to tools and techniques that are used while undertaking research to achieve the already set objective.

1.7.1. Modeling Technique

The recognizer in this study was built using the popular Hidden Markov Model (HMM). HMM is a powerful technique capable of robust modeling of speech. It is a parametric model that is particularly suitable for describing speech events. HMMs are also a succinct representation of speech events; therefore, they require less storage than many other strategies (Lee, 1989). It is most widely applied and said most successful speech modeling technique

1.7.2. Data Selection

In order to build a robust large vocabulary speaker independent speech recognizer, it is crucial to prepare phonetically balanced data sets, taking into account the many variations that can occur in speech. It follows that a training corpus should be sufficiently large, consisting of speech samples spoken by men and women from different age and if possible linguistic background.

Since the sound of a sub-word unit is also influenced by its context the speech data should include all phonemes in as many different phonetic contexts as possible. In practice, according to Wiggers,(2001) depending on the quality of the data and the complexity of the acoustic models, about ten- to thirty-thousand (phonetically rich) sentences are necessary. Recording such data set obviously is a major undertaking involving sub-tasks like selection of phonetically rich and phonetically balanced sentences, selection of appropriate participants, recording the data and the most time-consuming parts post-processing and transcribing the data. Studies of this kind in the developed languages are fortunate enough that many of these speech corpora are commercially available. However no such commercial database is there for Amharic. Despite this reality, Solomon ,(2005) has developed a corpus from 100 people with the required variety and each speaker having read around100 sentences, summing up to a total of around 10850sentences.This actually fulfills the requirements of the study.

Thus for this specific research, a total of 950 sentences read by 17 individuals were randomly selected of which 9 male and 8 female. 700 sentences read by 12 individuals were used for training purpose while 250 for testing.

1.7.3. Tool

HTK is a toolkit used in automatic speech recognition research and has been developed by the Speech Vision Robotics Group at the Cambridge University Engineering Department. It is written in C-programming language and works on UNIX and Windows platform Young et.ai.;(2002). (Most of the information discussed below is taken from HTK book).

In building HMM-based speech recognizers using HTK, the tools that may be used in the process can be categorized into 4 main classes; namely, data preparation tools, training tools, testing tools and analysis tools.

1.7.4. Data Preparation Tools

In the process of building HMM-based speech recognizers, a set of speech data files and their associated transcriptions are required. The tool HSLAB is used to record the speech data and to manually annotate it with any required transcriptions. HSLAB is the only tool in the HTK Package, which makes use of the graphics library HGRAF.

HSLAB is invoked by typing in the command line.

1.7.5. The Recording and Labeling Environment

Once the speech data is recorded, it must be converted into the appropriate parametric form. The tool HCOPY is used for this purpose. This program reads one or more data files to a designated output file converting the data into a parameterized form. By setting the appropriate configuration variables, all input files can be converted to parametric form as they are read in. Thus, copying each file in this manner performs the required encoding. Most of the operations performed by HTK assume that the speech data is divided into segments and each segment has a name or a label. The set of labels associated with a speech file constitute a transcription and each transcription is stored in a separate file.

These transcription files should be prepared. The tool HLED is an editor for manipulating label files. It works by reading in a list of editing commands from an edit script file and then makes an edited copy of one or more label files. This causes HLED to be applied to each labFile.The lab Files may be master label files in turn using the edit commands listed in cmdFile.

1.8. Training Tools

The next step of system building using HTK is to define the topology or layout of the Hidden Markov Models by writing a prototype definition. The purpose of the prototype definition is to specify the overall characteristics and blueprint of the HMM and the actual parameters is computed later by the training tools. Though sensible values for the transition probabilities must be given, the training process is very insensitive to these. An acceptable simple strategy for specifying sensible values for these probabilities is to make all of the transitions out of any state equally likely. Once the structure and overall form of a set of HMMs is defined, the next step is to estimate the parameters of the HMMs from examples of the data sequences. This process of parameter estimation is usually called training. The basic operation of the HTK training tools involves reading in a set of one or more HMM definitions, and estimate the parameters of these definitions using the speech data. The speech data are normally stored in parameterized form such as LPC or MFCC parameters. In order to perform the required parameter estimation, HTK supplies four basic tools; namely, HCOMPV, HINIT, HREST, and HEREST. Which tool to use depends on whether the HMM models to be used are whole-word based or sub-word based. To build whole-word based HMMs, the most common approach is to

calculate initial parameters for the model using HINIT and then use HREST to refine the parameters using Baum-Welch re-estimation algorithm.

For sub-word based models, a different approach called sub-word modeling is pursued.
Sub-word modeling refers to a technique whereby one HMM is constructed for each sub-word unit (Rabiner and Juang, 1993). The core process in sub-word modeling involves the embedded training tool HEREST. HEREST uses continuously spoken utterances as its source of training data and simultaneously re-estimates the complete set of sub-word HMMs. For each input utterance, HEREST needs a transcription i.e. a list of the sub-word units in that utterance. HEREST then joins together all of the sub-word HMMs corresponding to this transcription to make a single composite HMM for each utterance. HEREST is invoked via the command line. This causes the set of HMMs given in HMM list to be loaded. The given list of training files is then used to perform one re-estimation cycle. As always, the list of training files can be stored in a script file if required. On completion, HEREST outputs new updated versions of each HMM definition.

In order to use HEREST, it is first necessary to construct a file containing a list of all
HMMs in the model set with each model name being written on a separate line. The names of the models in this list must correspond to the labels used in the transcriptions and there must be a corresponding model for every distinct transcription label.
However, before training the recognizer using HEREST, some pre-processing is required. The sub-word models must be initialized and one initialization strategy is to make all models equal initially and move to embedded training. This is called flat-start training. The idea behind flat-

start training is to calculate the global data mean and covariance and assign these to all component means and covariance as start up values. This is done by the tool HCOMPV. HCOMPV is invoked via the command line: HCOMPV hmm train Files where hmm the name of the physical is HMM whose parameters are to be initialized. The effect of this command is to compute the covariance of the speech training data and then copy it into every Gaussian component of the given HMM definition.

1.9. Recognition Tools

HTK provides a single recognition tool called HVITE. HVITE is a general purpose word recognizer. It will match a speech file against a network of HMMs and Outputs a transcription for each of the recognized utterances. HVITE takes as input a network describing the allowable word sequences, a dictionary defining how each word is pronounced and a set of HMMs. It operates by converting the word network to a sub-word network and then attaching the appropriate HMM to the sub-word units. Recognition can then be performed on either a list of stored speech files or on direct audio input. The word networks needed to drive HVITE are usually either single word loops in which any word can follow any other word or they are directed graphs representing a finite-state task grammar.

1.10. Analysis Tool

Once the HMM-based recognizer has been built, it is necessary to evaluate its performance. This is usually done by using some pre-recorded test sentences to transcribe and match the recognizer

output with the correct reference transcriptions. This comparison is performed by a tool called HRESULTs which uses dynamic programming to align the two transcriptions and then count substitution, deletion and insertion errors. HRESULTS is invoked by typing the command line. When HRESULTS is invoked an output of the following form will appear:

----------Overall Results----------

SENT: %Correct =

WORD: %Correct=

==========

The first line gives the sentence-level accuracy based on the total number of label files which are identical to the transcription files. The second line is the word level accuracy based on the matches between the label files and the transcriptions. In the second line,

H is the number of correct labels (Hits), D is the number of deletions, S is the number of substitutions, I is the number of insertions and N is the total number of utterances.

.

1.11. Evaluation and Testing Techniques

The most important and the commonest testing parameter used in evaluating speech recognition systems is accuracy of recognition. According to (Zegaye, 2003), accuracy of the recognizer is the most important and common parameter used to evaluate speech recognition system .Thus the researcher used accuracy of the recognizer to evaluate the performance.

1.12. Thesis Organization

This paper is organized into five chapters. The first chapter gives background information about pronunciation variation and ASR. It also justifies the need for the study, presents the objectives

of the study and discusses the methodology followed throughout the work. The second chapter presents review of related literature. It also deliberates on the fundamentals of ASR and the various dimensions of ASR. The third chapter deals with the Amharic pronunciation variation and provides an overview of the language under study. Chapter 4 discusses experimentation and implementation details of the Amharic Speech Recognition System and the analysis made based on the findings. The last chapter presents the conclusions drawn and the recommendations made.

CHAPTER TWO.

ASR CONCEPTS AND REVIEW OF RELATED LITERATURE

2.1. ASR Concepts

Automatic speech recognition is the process of converting an acoustic signal to a set of textual words (Zue et al., 1996). It is one of the fastest developing fields in the framework of speech science and engineering. The first attempts (during the 1950s) to develop techniques in ASR, which were based on the direct conversion of speech signal into a sequence of phoneme-like units, failed. The first positive results of spoken word recognition came into existence in the 1970s, when general pattern matching techniques were introduced (Jackson, 2005).

Automatic speech recognition (ASR) is concerned with how users interact with their computers. Some users can interact with their computers using the traditional methods of a keyboard and mouse as the main input devices and the monitor as the main output device. Due to one reason or another some users cannot be able to interact with machines using a mouse and keyboard device (Jackson, 2005).

Speech recognition systems help users who are not able to use the traditional input and output devices. The list of applications of automatic speech recognition is so long and is growing; some of known applications include virtual reality, multimedia searches, auto-attendants, travel information and reservation, translators, and natural language understanding, to mention a few. Speech recognition system can be classified on the basis of the constraints under which they are

developed and which they consequently impose on their users. These constraints include: speaker dependence, type of utterance, size of vocabulary, linguistic constraints, and environment (Solomon, 2005).

2.1.1. Types of Speaker

A speech recognizer can be developed to recognize only read speech (discrete) or to allow the user speak spontaneously (continuous). Only read speech (Isolated) wordrecognition, in which each word is separated by pauses, is much easier than recognizing continuous speech, in which words run into each other and have to be segmented. Speech is said to be continuous when it is uttered as a continuous flow of sounds with no inherent separation between them.

2.1.2. Linguistics Constraints

Most of the present speech recognition systems are unable to reliably determine the identity of a speech input (phone or word) based on the speech signal alone. To improve reliability, linguistic constraints are put on a recognizer by using a language model and pronunciation dictionary. They capture syntactical and lexical constraints respectively. The more constrained the rule of a language in the recognizer, the less freedom of expression the user has in constructing spoken message (Rabiner, 1989).

2.1.3. Environment

Speech recognizer may require the speech to be clean from environmental noises, acoustic distortion and transmission channel distortion or they may ideally handle any of these problems. Current speech recognizers give better performance in carefully controlled environment. Their performance rapidly degrades when they are applied in noisy environment. This is due to that the noise can take the form of speech from other speakers, equipment sound, air conditioner, factory and from the speaker himself in the form of lips smacks, cough or sneezes (Rabiner, 1989).

2.2. Approaches of ASR

From the early days of speech technology, automatic speech recognition system has been using various approaches in order to achieve better performance. Accordingly, major approaches used in automatic speech recognition are the following Juang and Rabiner (1993).

2.2.1. Acoustic Phonetic Approach

The earliest approaches to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. This is the basis of the acoustic phonetic approach (Hemdal and Hughes, 1967) which postulates that there exist finite, distinctive phonetic units (phonemes) in spoken language and that these units are broadly characterized by a set of acoustics properties that are manifested in the speech signal over time. Even though, the acoustic properties of phonetic units are highly variable, both with speakers andwith neighboring sounds

(the so-called co articulation effect),it is assumed in the acoustic-phonetic approach that the rules governing the variability are straightforward and can bereadily learned by a machine. The first step in the acoustic phonetic approach is a spectral analysis of the speech combined with a feature detection that converts the spectral measurements to a set of features that describe the broad acoustic properties of the different phonetic units. The next step is a segmentation and labeling phase in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phoneme lattice characterization of the speech.

The last step in this approach attempts to determine a valid word (or string of words) from the phonetic label sequences produced by the segmentation to labeling. In the validation process, linguistic constraints on the task (i.e., the vocabulary, the syntax, and other semantic rules) are invokedin order to access the lexicon for word decoding based on thephoneme lattice. The acoustic phonetic approach has not beenwidely used in most commercial applications.According to Rabiner and Juang(1993) acoustic-phonetic approach is an interesting idea but it lacks success in practical speech recognition systems due to difficulty of decoding phonetic units into word string (Lexical access problem), the difficulty of getting reliable phoneme lattice for the lexical access stage, the requirement of extensive knowledge of the acoustic properties of phonetic units, the choice of features is made mostly based on ad-hoc considerations and the design of sound classifiers is also not optimal.

2.2.2. **Pattern Recognition Approach.**

Unlike acoustic-phonetic approach, this approach basically uses speech pattern directly without explicit feature determination (in the acoustic phonetic sense) and segmentation. The pattern-matching approach (Rabiner and Juang, 1993) involves two essential steps namely, pattern training and recognition of pattern via comparison. Theessential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm. A speech pattern representation can be in the form of a speech template or a statistical model (e.g., HMM) and can be applied to a sound (smaller than a word), a word, or a phrase.

Template based approaches match Unknown speech compared against a set of pre-recorded words (templates) in order to find the best match Rabiner et al.;(1979). This has the advantage of using perfectly accurate word models. But it has also the disadvantage that pre-recorded templates are fixed, so variations in speech can only be modeled by using many templates per word, which eventually becomes impractical. Dynamic time warping is such a typical approach (Tolba et al., 2001).In this approach, the templates usually consists of representative sequences of features vectors for corresponding words. The basic idea here is to align the utterance to each of the template words and then select the word or word sequence that contains the best. For each utterance, the distance between the template and the observed feature vectors are computed using some distance measure and these local distances are accumulated along each possible alignment

path. The lowest scoring path then identifies the optimal alignment for a word and the word template obtaining the lowest overall score depicts the recognized word or sequence of words.

Stochastic modeling entails the use of probabilistic models to deal with uncertain or incomplete information (Katti, 2009). The term stochastic refers to "the process of making a sequence of non-deterministic selections from among sets of alternatives" (Markowitz, 1996). It consists of employing a probabilistic model for the uncertainty or incomplete information that is inherent in speech signals. In speech recognition, uncertainty and incompleteness arise from any sources; for example, confusable sounds, speaker variability, contextual effects, and homophones words. Thus, stochastic models are particularly suitable approach to speech recognition.

Pattern recognition approach is the method of choice for speech recognition because of simplicity of use and robustness and invariance to different speech vocabularies, users, feature sets pattern comparison algorithms and decision rules. Because of this, pattern recognition approach is appropriate for different kinds of speech units, populations, background environment, and transmission condition. In addition, this approach provides high performance on any task that is reasonable for the technology Martha (2003).

2.2.3 **Artificial Intelligence Approach**

The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. In this, it exploits the ideas and concepts of Acoustic phonetic and pattern recognition methods. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram (Holmes and Holmes, 2001).

Some speech researchers developed recognition system that used acoustic phonetic knowledge to develop classification rules for speech sounds. While template based approaches have been very effective in the design of a variety of speech recognition systems; they provided little insight about human speech processing, thereby making error analysis and knowledge-based system enhancement difficult.

On the other hand, a large body of linguistic and phonetic literature provided insights and understanding to human speech processing. In its pure form, knowledge engineering design involves the direct and explicit incorporation of expert speech knowledge into a recognition system. This knowledge is usually derived from careful study of spectrograms and is incorporated using rules or procedures. However, this approach had only limited success, largely due to the difficulty in quantifying expert knowledge. Another difficult problem is the integration of many levels of human knowledge phonetics, phonotactics, lexical access, syntax, semantics and pragmatics. Lesser et al. (1975); Lippmann (1987) at Alternatively, combining independent and asynchronous knowledge sources optimally remains an unsolved problem. In more indirect forms, knowledge has also been used to guide the design of the models and algorithms of other techniques such as template matching and stochastic modeling. This form of knowledge application makes an important distinction between knowledge and algorithms. Algorithms enable us to solve problems. Knowledge enables the algorithms to work better. This form of knowledge based system enhancement has contributed considerably to the design of all successful strategies reported. It plays an important role in the selection of a suitable input representation, the definition of units of speech, or the design of the recognition algorithm itself (Katti, 2009).

2.3. Components of ASR

The underlying assumption behind any recognition system is that the waveform of a speech signal that comes out of a speaker's vocal apparatus is a realization of the concept that was in the form of symbols in his/her mind. When a source conceives an idea to speak out, it was understood symbolically. The moment it gets out to the channel, it materializes in the form of speech signals or sound waves. Thus, one direct and possible approach for a computer based speech recognition system to recognize an utterance is inferring the original symbols from the speech signals (Young et.al, 2002). Speech recognizer can be said to possess three components (Zegaye, 2003): feature extraction , language modeling and acoustic modeling and decoding.

2.3.1. Feature Extraction

Feature extraction refers to the first stage of ASR, whereby the input acoustic signal is converted to a sequence of acoustic feature vector. Ideally the method of feature extraction should preserve all the perceptually important information for making phonetic distinction, while not being sensitive to acoustic variations that are irrelevant phonetically. The main task of the feature extraction is to extract features from a speech signal. The aim is to sufficiently represent the characteristics of the speech signal with reduced redundancy (Chengand Abdulla, 2005). There are three major types of feature extraction commonly used in speech recognizers Wester,(2002). They are Linear Prediction Cepstral Coefficients (LPCC), Perceptual Linear Prediction Coefficients (PLP) and Mel Frequency Cepstral Coefficients (MFCC).

Linear Prediction Cepstral Coefficient (LPCC)

Linear Prediction Cepstral Coefficients (LPCC) has been commonly used in many speech recognition applications for many years. The notion behind LPCC is to model the human vocal tract by a digital all-pole filter and passes through the following.

A/ Pre-emphasis and Windowing

The first step of the algorithm is pre-emphasis. The idea of pre-emphasis is to spectrally flatten the speech signal and equalize the inherent spectral tilt in speech. Pre-emphasis is implemented by a first order FIR digital filter (Mantha, Duncan and Zhao, 2001).

B/ Linear Predictive Analysis

In human speech production, the shape of the vocal tract governs the nature of the sound being produced. In order to study the properties quantitatively, the vocal tract is modeled by a digital all-pole filter. The LPC lies on the assumption that the space between the vocal cords called glottis, produces speech signal which is characterized by its intensity (loudness) and frequency, which determines the speech of the sound. LPC analyzes the speech signal frames by estimating the formants, removing their effects from the speech signal intensity and frequency of the remaining buzz (Wiggers, 2001).

The basic intention of LPC is to determine the formants from the speech signal which is done by different equations called a linear predictor.

C/ Cepstral Analysis

Cepstral analysis refers to the process of finding the cepstrum of a speech sequence. Cepstrum, whose spelling is formed by shuffling the characters of the word spectrum, is a time-domain representation of a signal. Cepstrum is defined as the inverse Fourier transform of the logarithm

of a signal's spectrum. It has also been used to determine the fundamental frequency of human speech.

Perceptual Linear Prediction Coefficients (PLP)

Perceptual Linear Prediction (PLP) coefficient is another feature extraction technique, which tries to emulate the human auditory system.

Mel Frequency Cepstral Coefficient

Mel Frequency Cepstral Coefficients (MFCC) is one of the most commonly used feature extraction front-ends in speech recognition systems. It deals with power spectrum of speech signal which describe the frequency content of the signal over time. The purpose is to reduce the number of data characterizing the signal and shows a limited parameter or coefficient discriminating and robust. The technique is so-called FFT-based, which means that feature vectors are extracted from the frequency spectra of the windowed speech frames.

2.4. Issues in Pronunciation Variation Modeling

One of the main challenges in pronunciation modeling is to know which variation we are attempting to model. The effects of the acoustic models, the lexicon, and the language model will interact, even the choices at the speech pre-processing stage will influence the variation modeling (Amdal and Fosler –Lessier, 2002).

According to Wester(2002), there are two questions which cover most of the issues that must be addressed when modeling pronunciation variation:

1. How is the information obtained that is required to describe pronunciation variation?

2. How is this information incorporated in the ASR system?

2.2.4 Obtaining Information

Information about pronunciation variation can be acquired from the data itself or through (prior) knowledge; also termed the data-derived and the knowledge-based approaches to modeling pronunciation variation. One can classify approaches in which information is derived from phonological or phonetics knowledge and/or linguistic literature under knowledge-based approaches . In contrast, data-derived approaches include methods in which manual transcriptions of the training data are employed to obtain information or automatic transcriptions are used as the starting point for generating lists of variants.

Although the above approaches are useful, to a certain extent, for generating variants, they all have their own drawbacks too. The linguistic literature, including pronunciation dictionaries are not exhaustive; not all processes that occur in spontaneous speech (or even read speech) are described in the linguistic literature, or are present in pronunciation dictionaries. Furthermore, a knowledge-based approach runs the risk of suffering from discrepancies between theoretical pronunciations and phonetic reality (Cucchiarini, 1993).

2.2.5 Incorporating the Information in ASR

After the pronunciation variants are obtained, the next question that must be addressed is how the information should be incorporated into the ASR system. There are different levels at which this

problem can be addressed.In Strik and Cucchiarini (1999) a distinction was made among incorporating information on pronunciation variation in the lexicon, the acoustic models and the language models.

2.3 Modeling Pronunciation at the Level of Lexicon

With respect to the type of pronunciation variation to be modeled the choice is between variation within words and variation across word boundaries. This choice is influenced by factors such as the type of ASR and the language which is used, and the level at which modeling will take place. Modeling within-word variation is an obvious choice if the ASR makes use of a lexicon with word entries, because in this case variants can simply be added to the lexicon (Strik and Cucchiarini, 1999).

The lexicon typically consists of the orthography of words that occur in the training and theircorresponding phonetic transcriptions. During recognition, the phonetic transcriptions in the lexicon function as a constraint which defines the sequences of phonemes that are permitted to occur. The transcriptions can be obtained either manually or through grapheme-to-phoneme conversion. In pronunciation variation research one is usually confronted with two types of lexica: a canonical (or baseline) lexicon and a multiple pronunciation lexicon. A canonical lexicon contains the normative or standard transcriptions for the words; this is a single transcription per word. A multiple pronunciation lexicon contains more than one variant per word, for some or all of the words in the lexicon (Wester, 2002).

Modeling Pronunciation variation at this level needs adding of variants to the baseline recognition lexicon. In this way a lexicon is obtained that contains multiple pronunciations for some of the words. However, adding pronunciation variants to the lexicon usually also introduces new errors because the acoustic transcriptions of the added variants can be confusedwith those of other entries in the lexicon. This can be minimized by making an appropriate selection of the pronunciation variants (Amdal and Fosler –Lussier, 2002).

2.4 **Language Model**

Language modeling is a task of assigning a probability to a given sequence of words. It is crucial and indispensable for many speech and natural language applications such as automatic speech recognition (ASR), statistical machine translation (SMT) and optical character recognition (OCR) (Kim, 2004). It helps the applications especially in two ways. First, it reduces the search space of the problems. Provided by some probabilities estimated from a language model (LM), the search space can be effectively reduced by ignoring unlikely candidates, and thus the search problem becomes feasible. Most speech and natural language processing problems can be regarded as finding the most likely answer (word strings) given an input data (test set).

Second, language model actually improves an application's performance by providing contextual information. In ASR , it is practically impossible to distinguish **cent**, **sent**, and **scent** from one another unless some contextual information is given.ASR problem can be decomposed into two parts, the acoustic modeling problem, and the language modeling problem Katti (2009).

ˆW= argmax p/ (w)

P| (W) is Language Model

P| (A|W) is Acoustic Model

The ASR problem is to find the most likely word string W from the given acoustic evidence (input data) Katti (2009).

$\grave{W}$= argmax $\frac{P(W/A)}{W}$

By applying Bayes' formula equation can be rewritten as

P (W|A) = $\frac{(p(w)p(A|w))}{p(A)}$ and since A is fixed acoustic evidence is already given and doesn't change over the recognition process.Small vocabulary recognition system do not rely on language model to accomplish their tasks because they are mainly used in command and control signal that the vocabulary has to respond (Zegaye, 2003).A large vocabulary speech recognition system generally depends on linguistic knowledge. Hence, incorporation of knowledge of the language, in the form of a language model is essential for large vocabulary system.

2.5 **Acoustic Model**

Speech is in essence just a sequence of different sounds. Our brains are tuned to classify these sounds into basic phonetic units, or phonemes. From a sequence of phonemes we can distinguish words. From a pattern recognition point of view, this is quite an astonishing feat considering that the brain is also able to comprehend speech produced in different environments and by different speakers. Devising an algorithm for a computer to do the same is not a trivial matter.

An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes up a word. Each of these statistical representations is assigned a label called a

phoneme Mesfin,(2008). The core of an acoustic model lies in the capability of the feature vector to capture the distinctive property of the speech.

Speech recognition engines require two types of files to recognize speech Odelly (1995) They require an acoustic model, which is created by taking audio recordings of speech and their transcriptions (taken from a speech corpus), and compiling them into a statistical representations of the sounds that make up each word through a process called training. They also require a language model or grammar file, a file containing the probabilities of sequences of words.Once the signal has been transformed in to a parameterized form, it must be recognized or decoded and turned in to the underlying sequence of symbols (Odelly 1995). This decoding process requires patters against which unknown utterances can be compared.

The acoustic model, P (O/W), provides the probability that the speech data was observed for a given word sequences. The required probability distribution could be found by obtaining many examples of each word W and collecting the statistics of the corresponding vector sequences (Young, 1996). However, this is impractical for large vocabulary system instead word sequence is decomposed in to phonemes. Probability for word sequence is generated as a product of the acoustic and language model probability. The process of combining these two probability scores and sorting through all plausible hypotheses to select the one with maximum probability is called decoding or search (Ganapathiraju, 2002)

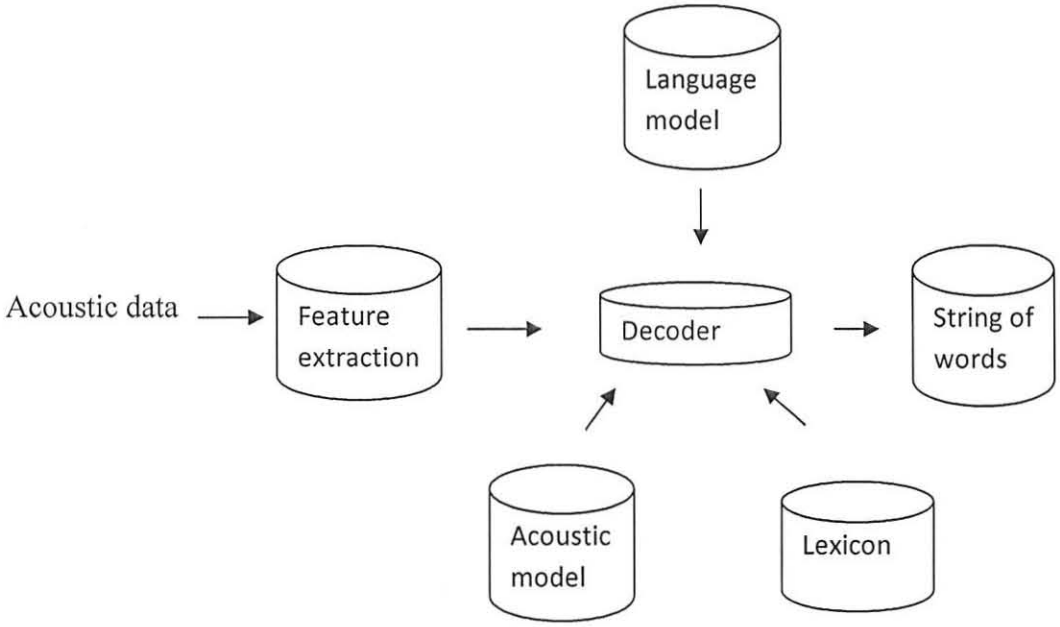

Fig.2.2. Basic Components of Speech Recognizer System (Holmes and Holme, 2001)

2.6 **Review of Related Literature**

Literature review was performed to investigate the underlying principles/theories of the various approaches, techniques and tools that were employed in the research. In addition, literature on the Amharic, especially those dealing with the phonetic/triphone features were reviewed. Moreover, to learn what others have done in the area and to better understand the problem, a comprehensive investigation of available empirical literature on automatic speech recognition was conducted. This section presents research done on ASR and pronunciation variation in ASR for Amharic.

Kinfe (2002) undertook research on Amharic entitled 'Sub-Word Based Amharic Speech Recognizer': An Experiment Using Hidden Markov Model. 170 word vocabulary was constructed and both speaker-dependent and speaker-independent. Models were built for 15 speakers (8 male and 7 female) in the age range of 20 to 30 using phonemes and tied-state tri-phones as the basic units of recognition. Five untrained speakers who had no involvement in training the models were also used to test the speaker-independent models. The results obtained are promising and have shown the potential of tied-state tri-phones as good sub-word units for Amharic. In fact, phonemes also have produced encouraging recognition performance.

Zegaye (2003) conducted research on 'Large Vocabulary, Speaker Independent Continuous Speech Recognition' for Amharic using Hidden Markov Model approach and Hidden Markov Model Toolkit. After the monophones speech recognition was built, it was promoted to tri-phone based system in which the right and left context were considered for optimal output.

Martha (2003) conducted research on 'Application of Amharic Speech Recognition System to Command and Control Computer: An Experiment with Microsoft Word.'
The study explored the possibility of developing Amharic speech input interface to command and control Microsoft Word. Fixed variance and variable variance based models HMM-based, speaker independent, small vocabulary, isolated Amharic word recognizers were developed. The performance of these recognizers was tested using the test data. Although both of them recognized all the test data correctly, the performance of recognizer with variable variance performed better than the recognizer with fixed variance in live environment. Thus, the

recognizer with variable variance was further considered for the development of the prototype Amharic speech input interface system.

Similarly, Solomon (2005) conducted research on Speech Recognition for Amharic. Hidden Markov Model with five optimal topology of CV syllable and a model of five emitting states and three emitting states with twelve Gaussian mixtures without skip and jumps were used; where the three emitting sates showed better performance.

When we come to our target study, pronunciation variation, Solomon (2005) conducted research on 'Multiple Pronunciation Model for Amharic Speech Recognition System.' The research has tried to show the pattern variations of sound units in Amharic for multiple pron unciation models. This is variation of sound units at lexical level due to dialects. After that an attempt to build a pronunciation dictionary for Automatic Speech Recognition (ASR) was made.

Tewodaj (2013) has undertaken research entitled " pronunciation variation in Amharic speech recognition with data- driven approach" the performance of the recognizer was 60.54% sentence accuracy and 73.23% word accuracy for canonical and 62.68% Sentence accuracy and 74.57% word accuracy for data driven (alternative)dictionary.

Similarly Solomon (2013) has conducted research on " comparing data- driven and knowledge based approach performance pronunciation variation modeling of Amharic ASR ". In the knowledge based performance pronunciation, 12.79% and 59.34% sentences were correctly recognized with 56.10% word accuracy. In the performance of data driven, pronunciation 13.13%

and 60.34% sentences were correctly recognized with 56.10% word accuracy and shows that data-driven has better performance than knowledge base.

The reviewed literature reveals that works done on the area of pronunciation variation handling were few in number. Thus further study on this area is mandatory to obtain better performance for Amharic ASR system.

CHAPTER THREE

PRONUNCIATION VARIATION IN AMHARIC

3.1. Amharic Language

Amharic is the working language of Ethiopia. As well as the official language of the Amhara, the South Nations, Nationalities and Peoples, the Gambela and the Benishangual Gumuz regional states and the city administrations of Addis Ababa and Diredawa.. Within the Semitic language family, it has the greatest number of speakers after Arabic. Amharic is one of the languages which have their own writing system cited by (Solomon, 2005). While the sound and the spelling systems of many languages do not correspond exactly (Clark et al., 1985), the degree of misfit between sound and spelling in Amharic is not significant. Provided that one knows the alphabets of the language, the chances of pronouncing a word upon seeing it written for the first time is quite simple; furthermore, upon hearing a word in Amharic, one can easily spell it correctly. There is disagreement among scholars whether the Amharic writing system is syllabic or alphabetic. Writing systems in which one symbol represents one syllable are called syllabic while writing systems in which one symbol represents one sound segment are called alphabetic (Clark et al., 1985). Some writers (Bender, 1976; Cowley, 1976; Mullen, 1986) argue that the Amharic writing system is syllabic while others (Tadesse, 1994; Baye, 1997) say that it is not syllabic. Amharic adopts the principle of writing one CV-syllable with one character. This reduces the number of signs to manageable proportions. I support Tadesse and Baye for this research because of its clearity.

Linguists decompose a spoken language into elements of linguistically distinct sounds called phonemes. These phonemes are determined and classified according to their corresponding articulatory configurations (Juang and Furui, 2000). Articulatory phonetics deals with how the human vocal apparatus is manipulated to produce sounds (Clark et al., 1985). The basic assumption of articulatory phonetics is that sounds are best described in terms of the configurations of the vocal tract necessary to utter the sounds.Sounds can also be classified as vowels and consonants. Vowel and consonant sounds are produced in fundamentally different ways. While consonants are articulated with a substantial degree of obstruction in the oral cavity, vowels are produced with a relatively free airflow (Clark et al., 1985).

3.1.2. Description of Amharic Phonemes

3.1.2.1.Amharic Consonant Phonemes

Of the 31 Amharic consonant phonemes, /p', p, s'/ are found in words that are borrowed from other languages. /p'/ is found in Greek-derived words like p'ap'as, t'ərəp'eza, p'et'ros. /p/ occurs in "European" words like police, pasta, parlama, politika, pant, ampol. The consonant sound /s'/ tends to occur in words borrowed from Ge'ez. Those people who do not have knowledge of Ge'ez and some people who live in rural areas tend to replace it with /t'/; for example, they may say /t'ələt /, /t'əbəl/, /t'əhay/ instead of /s'ələt /, /s'əbəl/, /s'əhay/ respectively (Getahun 1997:7, Baye 1994, Mulugeta 2001:9).

In this study the table produced by Getahun Amare (1997) have been adopted, because of its clarity and brevity.

Manner/place of Articulation	Bilabial	Labio-dental	Alveolar	Palatal	Velar	Glottal
Stop/plosives Vl	p(ፐ)		t(ት)		k(ከ)	ʔ(አ)
Vd	b(በ)		d(ድ)		g(ገ)	
Ejective	p'(ጰ)		t'(ጥ)		k'(ቀ)	
Fricatives Vl		f(ፍ)	s(ስ)	ʃ(ሽ)	$k^{w(i)}$	h(ሀ)
Vd			z(ዝ)	ʒ(ዥ)	$g^{w(\tilde{O})}$	
Ejective			s'(ጽ)		$k'^{w(n)}$	
Affricates Vl				ʧ(c)		
Vd				ʤ(ጅ)		
Ejective				ʧ'(C)		
Nasals	m(ም)		n(ን)	ɲ(ኝ)		
Lateral			l(ል)			
Trill			r(ር)			
Semivowels	w(ው)			y(ይ)		

Table 3.1: Amharic Consonant Phoneme Chart (Adapted from Getahun Amare 1997:5)

Amharic consonants are often described as including other, more marginal consonant sounds like the labialization of velars and glottals /g^w, k^w, k'^w, h^w/ and of other sounds, especially labials /b^w, f^w, m^w, p'^w, t'^w / (Baye 1994; 2008, Lulseged 1981:37, Hayward & Hayward 1999:45).

3.1.2.2. Simple Vowels

Amharic has seven simple vowel sounds, like other Ethio-Semitic languages. These vowel sounds are /ə/, /u/, /i/, /a/, /e/, /ɨ/, & /o/. These vowel sounds occur already in Ge'ez as the seven orders of the fidəl, as Cə Cu Ci Ca Ce Cɨ Co (Leslau 1995, Bender 1976).

Vowels can be described in terms of the height of the tongue (high, mid and low), the horizontal position of the tongue (front, central and back) and the condition of the lips (rounded and unrounded).

Getahun (1997) puts the seven Amharic vowel phonemes in the following way:

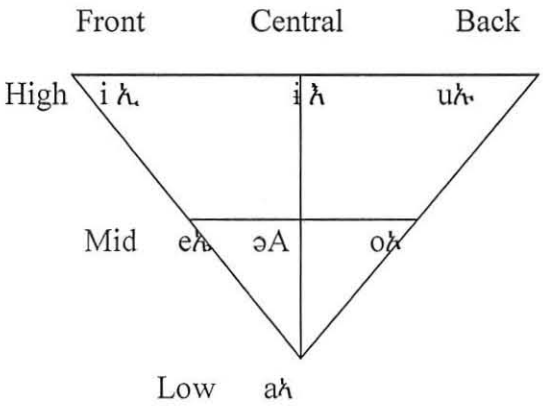

Fig 3.1:Amharic Vowel Phoneme Chart (Getahun 1997:14)

3.1.2.3. Amharic Orthography

The Ethiopic alphabet (commonlly referred to as a syllabary) was adapted from Ge'ez and is used to write Amharic and several other regional languages including Tigirgya and Tigre. In a syllabary, characters usually represent consonant-vowel cluster. In the first category of the Amharic writing system, called fidel, there are thirty-three core orthographic symbols, each of which has seven different shapes, usually known as orders, to represent the seven vowels, making the total symbols 231 (33x7). Each consonant and the seven vowels in combination represent CV syllables. Each of these consonant and vowel grapheme can appear independently or can form a combinant letter. Each consonant can form CV pattern except with epenthesis

vowel. The second category (4*5=20) consists of four labio-velar symbols, which have five orders the eighteen labialized consonant, which have only one order, are the third category. The fourth category is the representation of numbers from 1 to 10 and multiples of 10 each with different symbols. As with roman script characters run from left to right across the page.

There are some problems in the representation of some sounds in the orthography, allowing room for mispronunciation. However, the degree of misfit between sound and spelling in Amharic is not significant for the native speaker. Provided that one knows the alphabets of the Amharic language, the chances of pronouncing a word upon seeing it written for the first time are pretty good. Furthermore, upon hearing a word in Amharic, one has a good chance of spelling it correctly (Clark et al., 1985).

Nevertheless, in relation to the writing system, Bender (1976) has mentioned three problems:

A. Presence of several symbols (fidel) - like for the sound /ha/ .This means they have the same sound but different orthography.

B. Many of the syllabic characters are irregular and no standard system of handwriting is there.

C. There is no marking for the gemination of Amharic consonants that is why there is difficulty for the machine to recognize them.

There is also an ambiguity between six-order symbols with and without vowel and multiple representation of one sound by two or more letters, for example, ሰ, ሠ and ጸ, ፀ Due to these Amharic orthography variations, it is difficult for the machine to recognize them. The other

problem is possible representation of orthography through different spelling like ቋንቋ = ቁዋንቁዋ, which have the same meaning but they have the different orthography representation.

3.1.2.4. Gemination

Gemination in Amharic is one of the most distinctive characteristics of the regular rise and fall of the voice of the speech and also carries a very heavy semantic and syntactic functional weight similar to other languages. Gemination in Amharic share many common features (particularly with Semitic languages) but unlike many other languages gemination in Amharic is not shown in orthography. However, since there are relatively few minimal pairs , Amharic readers seem not to find this to be a problem. But when we come to ASR, this is a big problem because the machine doesn't identify the geminated forms with the particular meanings.

That is, consonant length can distinguish words from one another; for example,

Ala/ኦለ /"he said" as compared to A*lla/ኦለ*/ -" existed"

wAnna/ዋና/ - "chief" as compared to wAna/ዋና/- swimming .

gannA/ ገና/ - "chrismas" as compared to ganA/ ገና/ -"still "

Gemination is at times optional. For example አትሰብርም [AtsAbrIm]- you will not break it says also at the same time we can omit ት/tI/ and says አሰብርም [AsAbIrImI] this doesn't change the meaning both has the same meaning so it can be geminated or not geminated.

3.2. Pronunciation variation

Pronunciation variation refers to the fact that words can be pronounced in many different ways. Differences exist in the way speech is pronounced by various speakers, but even if the same speaker utters a word more than once, it will never be pronounced in exactly the same way. Humans usually have no difficulties in processing different pronunciation variants of the same word, since they have knowledge of pronunciation variation.

However, for machines, pronunciation variation creates a problem, because, in general, machines do not explicitly take into account the different ways in which words can be pronounced (Kessens, 2002).

The objective of automatic speech recognition (ASR) is to recognize what a person has said, i.e., to derive the string of spoken words from an acoustic signal. pronunciation variation impedes this process. Obviously, the variation is inter and intra speaker which is caused due to different factors, such as age, gender, dialect, emotion, speaking style etc. this is true for Amharic. This variation affects the performance of ASR. Due to the above described variation the objective of ASR becomes more difficult to achieve, as the pronunciation variation may lead to recognition errors.

3.3. Sources of Pronunciation Variation

3.3.1. Intra-Pronunciation Variation

Intra-speaker variation refers to pronunciation variation of the same speaker (Laver, 1994). The variation can be caused by speaking style, speaking rate and co-articulation. Speaking styles

refers to as stylistic variation; this type of variation depends on whether the speech is scripted, planned or spontaneous. It has been shown that speaking rate can have a dramatic impact on the degree of variation in pronunciation. Co-articulation, which is the overlapping of adjacent articulations, affects the way words are pronounced and variation in the degree of co articulation causes pronunciation variation and Supra segmental features: for instance, word stress, sentence stress, intonation, frequency of occurrence of a word, position of a word in a sentence, and position of a consonant or vowel within a syllable all affect the pronunciation of word (Wester, 2002).

3.3.2. Inter-Speaker Pronunciation Variation

Inter-speaker variation is caused by anatomical differences between speakers. It also exist due to the fact that speaker of the same language may speak different dialect or speak with a different accent. The accent will depend on factors such as, region of origin, socio economic back ground and level of education, sex and age (Judith, Kessens, & Strik, 2002). For example, male and female speakers and children have different speech characteristics (Wester, 2002).

Age

Every human being goes through the process of ageing. This is a very complex process, which affects the way we speak our voices and speech patterns change from early childhood to old age. Although most changes occur in childhood and puberty, age-related variation can be observed throughout our adult lives into old age (Schotz, 2007). Children have shorter vocal tract and

vocal folds compared with adults. This results in higher of formants and fundamental frequencies. The high fundamental frequency is reflected as a large distance between the harmonics, resulting in poor spectral resolution of voiced sounds. Thus, the difference in vocal tract size, results in pronunciation variation (Benzeguiba, et al., 2006).

Voice quality

Biemans (2000) states, voice quality refers to laryngeal qualities or a specific phonation type, and sometimes it is used in a broad sense as the total vocal image of a speaker, including for instance pitch, loudness, and phonation types. In the present study, 'voice quality' is used in the latter sense. Voice features can be short-term, medium-term, or long term. These three time domains have different functions attached to them short-term convey meaning through the sequential ordering of phonological and grammatical units in larger structures, i.e. consonants, vowels, words, and longer utterances. Medium-term conveys the emotional state of the speaker (e.g. Anger, happiness, disappointment).

Gender

Mean fundamental frequency, which is associated with the perceptual notion of pitch, is commonly considered as the major difference between adult male and female voices. Mean F0 would be around 120 Hz for men and 200 Hz for women but these values slightly vary through age and are broadly lower for smokers Ladefoged (2001). Phonation type also seems to depend on the speaker's gender. Female voices are often considered more breathy than male voices. Vowel formants of female speakers tend to be located at higher frequencies as well as consonant noise. Vocal folds become longer and thicker in male speakers, and this explains why they tend

to vibrate more slowly than those of women. Another important anatomical issue is vocal tract length, that is, the distance from the vocal folds to the lips. The average length of the adult female vocal tract is about 14.5 cm, while the average male vocal tract is 17 to 18 cm long (Erwan, 2012).

Traunmüller (1989), States that F0 are 120 Hz for men and 210 Hz for women. The mean values change slightly with age. For men, the decrease in FO that is most dramatic during puberty has been observed to continue with successive deceleration until about 35 years of age at about 55 years of age, FO begins to rise again. For women, FO is stationary up to the age of menopause, when it decreases to reach a minimum that is about 15 Hz lower around 70 years of age Liberman (1977).

Accent

Accent refers to the way in which a speaker pronounces, and therefore refers to a variety which is phonetically and/or phonologically different from other varieties Chmbers and Peter(2004). intra speaker dependent accent brings variation in pronunciation and it may depend on different factors. Judith, Kassens & Strik (2002) say that the accent will depend on factors such as region of origin, socioeconomic background, level of education, sex and age. This is true for Amharic. Because of accent different speaker utter the same word differently. The speakers 'accents are categorized according to their geographical affiliation (Huckvale & Tjalve, 2005).

Dialect

There is variation in language, a variation that also concerns the lexical units. In any language, one can expect to come across instances where certain speech differences may exist between various groups of people. Speech differences in a language may arise due to the influence of a language in an adjoining area. Sometimes they are confined to the use of different sounds or tones, in order to express the same thing.(Laver,1994). These speech differences are called variants. Variants of the same language are called dialect. Any language spoken by more than a handful of people exhibits this tendency to split into dialects, which may differ from one another along many dimensions of language content, and function: vocabulary, pronunciation, grammar, usage, social function, artistic and literary expression. When the language of one group of people shows regular variations from that used by other groups of speakers of that language, we speak of a dialect Munzhedzi & Mafela (2008). Dialect is the very big issue for pronunciation variation in Amharic while other constraints have also their own contribution.

3.4. Approaches for Representing Pronunciation Variation

Approaches of pronunciation variation can be roughly divided into pronunciation variants being either derived from a corpus of pronunciation data or from pre-specified phonological rules based on linguistic knowledge (Strik & Cucchiarini, 1999).

3.4.1. Knowledge Based Approach

In this approach information about pronunciation is derived from knowledge sources, such as pronunciation dictionaries, phonological rules hand-crafted by linguistic experts or extracted

from the literature (Wester, 2002).This approach makes use of existing knowledge that was derived by experts. This can be dictionaries or results from linguistics studies on pronunciation variation. The gathered information is often used to derive rules that are able to generate typical pronunciation variants from canonical pronunciations or from the orthography of a word. The pronunciation variants that are generated by applying the rules can then be added to the dictionary.

For Amharic there is no dictionary prepared with multiple pronunciations. Thus, to use this approach it obligatory to know each word pronunciation probability by the same speaker or different speakers to prepare dictionary or consult experts. The advantage of this approach is that it is completely task independent, since it uses general linguistic and phonetic rules and can thus be used across corpora and especially for new words that are introduced to the system.

The drawback however is that the rules are often very general and thus many variants are generated, some of which might not be observed very often. On the other hand the existing knowledge might not cover all aspects that are needed for the current task and thus not enough variants might be generated. Furthermore no information on how often the generated variants appear in the data under consideration is given.

3.4.2. Data Driven Approach

Data-driven approaches try to derive the pronunciation variants directly from the speech signal. This can help avoiding over-generation since only variants that really occur in the data are used. It furthermore allows the computation of application likelihoods. On the other hand this method

is very much database-dependent and variants that occur frequently in one speech corpus do not necessarily occur frequently in other corpora.

No rules or hand-labeled data were used, only a baseline recognizer. This recognizer was used to make an *N*-best list of pronunciations without any prior knowledge of the vocabulary other than the number of words and the boundaries of each word, which is usually present in an orthographic transcription (Wester, 2002).

The first step in rule generation is finding an alternative transcription that can reveal the true pronunciations of the speakers. We can also use knowledge as a starting point: if we have hand labeled data, pronunciation rules can be derived from comparing this transcription with the reference. Usually only a word transcription exists, and if the reference lexicon contains several pronunciations, the recognizer is used to choose pronunciation by forced alignment.

CHAPTER FOUR

EXPERIMENTATION

In this study an attempt is made to design and construct the speech recognizer for Amharic that is capable of recognizing Amharic speech (sounds). The intention was developing a system that is capable of handling variation that exists in Amharic during ASR development. To meet these requirements the recognizer was designed to recognize large vocabulary, continuous speech and is speaker independent. This was implemented using phonemes as base unit. The discussion of the experiment is presented in accordance to the steps that should be followed while building such speech recognizer.

Accordingly, the chapter is organized in to four parts: *preprocessing, Training, Optimization (Enhancement)* and *Evaluation.*The preprocessing part presents the way the data preparation task was performed, the developed language model and the prepared dictionary. The training portion concerns how the acoustic models were developed, the steps followed in training and building the mono-phone level recognizer, and the tools with their accompanying scripts used in the process. It also discusses the mechanisms followed in refining and optimizing the mono-phone recognizer. The final part is on the testing and performance evaluation of the systems.

4.1. Preprocessing

4.1.1. Dictionary Preparation

Dictionary contains words with its pronunciation. The pronunciation can be canonical, knowledge based, data driven or hybrid. Canonical dictionary contains standard pronunciation of the language under study. It has only one pronunciation per word. Knowledge based is a dictionary where the information about pronunciation variant is derived from existing knowledge that was derived by experts. In case of data driven the pronunciation variant is derived from the speech signal. The other way of dictionary preparation is using hybrid. This technique combines knowledge based and data driven to get pronunciation variants for a specific word.

No external and standardized pronunciation dictionary has been prepared for Amharic. Thus for the canonical dictionary a python program was developed to create word lists with their associated pronunciation from the text data base. After canonical dictionary was generated automatically, alternate dictionaries using knowledge based (100 words), data driven (100 words) and hybrid 100 words; 50 words using knowledge based and 50 words using data driven were developed by adding variants to the already prepared dictionary manually. Then using the tool HDman in HTK and the produced dictionary, two different new dictionaries were created: one with a short pause at the end of every pronunciation and the other without it.

4.1.2. The phoneme Sets Extraction

Phoneme set extraction is one of the important tasks to be performed while developing the model. Accordingly, from the dictionary prepared 37 sets of phones were extracted. The silent

phoneme was added to the phoneme list so as to handle the silence filler dictionary that causes error during training. The first, silent models longer periods of silence. These silences occur between sentences or when a person is not speaking at all. The second phoneme, short pause, also represents silence, but only periods of short duration. This is the kind of silence that occurs between words. The former represents periods of pure silence that can greatly vary in length. It usually demarcates sentence boundaries. The later represents optional periods of silence shorter than the duration of a phoneme, and are usually found between words.

4.1.3. File Transcription

Speech is divided into segments and each segment should have a name or a label. The set of labels associated with a speech file constitute a transcription. Two types of transcription files were prepared out of the utterances: word level and phone level transcriptions. This was because, as the system is phoneme based, later during training examples of the phones were needed to adjust the parameters of these models. That is, to train a set of HMMs every file of training data must have an associated phone level transcription. This file transcription is derived from the already developed dictionary. Thus, the file transcription was done according to the dictionary developed for each model. To this end, the file transcription developed for different model is different with respect to the dictionary.

These phone level transcriptions were created from the word level transcriptions. A program was executed to prepare the word level transcription first. It reads each word from the transcription text and put them in a separate line. Then a file called Master Label File (MLF) containing all the

transcriptions of the words in the utterances, indexed by the file name of the utterances was produced. All the utterances had unique file names and the file was prepared in a label file format required by HTK. Then using the already created pronunciation dictionary and the above phone level MLF, the phone level transcription was created by the tool HLed in HTK.

4.1.4. Feature Vector Extraction

A raw audio signal received from a microphone, is too complex to deal with when it comes to the task of speech recognition. It needs to be converted into a more manageable form. This is the primary role of the feature extraction .Feature extraction can be understood as a step to reduce the dimensionality of the input data, a reduction which inevitably leads to some information loss. Literature reveals that Mel Frequency Cepstral Coefficients (MFCC) is one of the most commonly used feature extraction front-ends in speech recognition systems. It deals with power spectrum of speech signal which describe the frequency content of the signal over time.

In ASR development acoustic models are created from the already recorded speech and its transcription which is then processed to create statistical representation of the sound that make up every word. The acoustic model is later used by the speech recognition decoding engine to actually recognize the speech. Therefore, creating feature vector is essential task for creating the acoustic model.In HTK, HCopy is a tool used for translating audio files to feature vector files using Mel scale cepstral coefficients. The feature file was created using the following script:

HCopy -A -D -T 1 -C wav config.config –S code train.scp.

4.1.5. Language Model

Language model is a crucial part of a speech recognition system. It tells the system how likely it is that a certain string of words is uttered. By so doing, it imposes a grammar onto the system. Language model is the glue that holds together the set of acoustic models in the word network that is ultimately used for recognition. The Language Model used in this work was a backed-off bigram language model. It wasdeveloped using two of the HTK tools HLstats and HBuild. HLstats was primarily used to generate the bigram probability matrix. It reads in the sorted word list and the word level MLF.Using the word level transcriptions and statistics on the number of occurrences of eachword and each combination of two words, the probability matrix was prepared. These statistics were then used to create *backed-off bigram language*models for the training, testing and evaluation sets, using the HBuild tool which translated the gathered statistics into HTK Standard Lattice Format, that are used for storing word models andmultiple hypotheses from the output of a speech recognizer. This was done using HBuild and HLstats.

4.2. Training the Acoustic Model

In the previous activities all the required resources for training the recognizer and developing the models were prepared. Here the tasks performed in developing the recognizer and the techniques followed in refining and improving performances is explained.

4.2.1. HMM Prototype

The very first step in training HMMs is the selection of a Hidden Markov Model topology for the acoustic models. This was done by defining a prototype model. Since a phoneme based recognizer was built a model represented a phoneme. The parameters of the model are not important because they are to be modified later during training. However it helps to define the models. Thus the means and variances of all the states in the model were simply assigned a value of 0 and 1 respectively.

The topology of the model consisted start and end states and three emitting states, using single Gaussian density functions. The states were connected in a left-to-right way, with no skip transitions.A five state topology was chosen depending on reviewed literature for it resulted in better performance. Single mixture was used first and later the mixture was incremented.

4.2.2. Initial models

Literature reveals that training of HMMs can be done using the Baum-Welch algorithmused to find the unknown parameters of hidden markov model (HMM). It makes use of the forward-back ward algorithm. But to ensure proper and fast convergence of the models ,sensible initial values have to be calculated for the transitions parameters and the means and variances of the state density functions before the Baum-Welch algorithm can be used. HMMs are pretty sensitive to initial values so this stage is crucial for the entire process. Good initial models can be obtained by assuming an HMM as a generator of speech vectors. The training examples of the phones corresponding to the model whose parameters are to be estimated can be viewed as the output of

this model. Thus if the state that generated each vector in the training data was known, then the unknown means and variances could be estimated by averaging all the vectors associated with each state (Young et.al., 2002).

In HTK this concept is implemented using the HCompv tool. Based on the above prototype model, the tool HCompv was executed to create another average model.

This produced a new prototype version and stored it in the directory hmm0. It scanned the set of training data files and compute the global mean and variance and set all of the Gausians in a given HMM to have the same mean and variance.The –f option was used to create a variance floor macro, called vfloor, which is equal to0.01 times the global variance and is used to set the floor on the variance estimated in the subsequent steps.Finally, a Master Macro File (MMF) which contained the initial model set by the namehmmdefs was created. This was done by manually copying the prototype for each phone and renaming it accordingly.

4.2.3. Embedded Re-estimation

Once the initial monophone model set was available, the next task was re-estimating the model parameters. And an embedded re-estimation strategy was used, that simultaneously updated all of the HMMs in the system using all of the training data.

In embedded training all models are trained in parallel rather than individually. That is each utterance is processed in turn and the accompanying transcriptions are used to construct a composite HMM which spans the whole utterance. This composite HMM is

made and implemented by concatenating instances of the phone HMMs corresponding to each label in the transcription. The Forward-Backward algorithm is then applied and the sums needed to form the weighted averages are accumulated. When all of the training files have been processed, the new parameter estimates are formed from the weighted sums and the updated HMM set is created. This embedded procedure is implemented in the HTK tool *HERest* which performs exactly one iteration of the algorithm each time it is run.

4.2.4. Fixing the silence models

In the previous step a three states left-to-right HMM was generated for each phone and for the silence model. Though the sil model had the same topology as the other phones, it is supposed to take care of periods of silences that can vary greatly in length, from a few milliseconds up to a few seconds. Thus for this model to handle such situations, transitions added from its second state to the fourth and from the fourth state back to the second. This was done to make the model more robust by allowing individual states to absorb the various impulsive noises in the training data. The backward skip allows this to happen without committing the model to transit to the following word. So far, training was done without the short pause model .But from this on, it was introduced using another monophones, the phone set with sp, which was generated using the tool HDman and the phone level MLF. This MLF was made to contain sp at the end of each word while it was generated using hled; out of a reasonable assumption that there will always be such unnoticed pause there in. Otherwise, the transcription of each utterance had to be prepared in such a way that the sp was transcribed whenever it is observed in all the recorded speeches.

The model was designed to have three states, i.e., only one emitting state in the middle and two non-emitting states at left and right. Then the emitting state was made to be tied to the center state of the sil model resulting in the so called tee model.

4.3. Refinements and Optimization

Literatures indicate that monophone systems have a number of short comings and usually came up with poor performance. This basically is attributed to the fact that they are assumed to be insensitive to context variations. Thus they should be refined and strengthened. Accordingly, two of the commonest refining and optimization techniques were attempted in this experiment: mixture component splitting and promoting the monophones to tied state triphones.

4.3.1. Multiple Mixtures

Multiple mixture systems are said to improve recognition results considerably, because they help avoid the problem resulting from the usage of the same type of distribution for different models and different states. If an HMM state is made to contain multiple mixture components, then the training vectors would be associated with highest likelihood mixture component. The number of vectors associated with each component within a state can then be used to estimate the mixture weights. So the best strategy employed in many systems is incrementing the mixture components in stages by a factor of N.

Thus taking the single Gausian monophone system above, the mixtures were incremented by a factor of two until 8 mixture component HMMs were obtained. Then at each stage re-estimating and checking recognition results was performed.In HTK, the conversion from single Gaussian HMMs to multiple mixture component HMMs is implemented using the HHEd MU command which will increase the number of components in a mixture by a process called mixture splitting. In this method the command works by repeatedly splitting the mixture with the largest mixture weight until the required number of components is obtained.

4.3.2. Tied State triphones

The phone models described so far were context independent. To capture these effects, called co-articulations, models are needed that take into account the context of a phone. One way of modeling co-articulation effects is using triphones. Triphones model the context by taking in to consideration the left and right neighboring phones. If two phones have the same identity but different left or right context they are considered as different triphones.The triphone models constructed here were word internal because the other type, cross word triphone, requires far more data though they are powerful. Using HTK, the context dependent triphones were prepared by simply cloning monophones and then re-estimating using triphones transcriptions. First triphones' transcription was needed to train the triphone system. This was prepared using the HLEd command:

*HLEd –n triphones1 –l * -i wintri.mlf mktri.led phones1.mlf.* This converts the monophones transcriptions in phones1.mlf to an equivalent set of triphone leveltranscriptions in to wintry.mlf.

At the same time, a list of triphones was written to the file triphones1. The WB commands defined sp and sil as word boundary symbols i.e. the start and end phones in a word.

The cloning of models was then performed by executing the HHEd command.The file mktri.hed was generated using a perl script. Then, two more re-estimation was performed, putting the model set in hmm10 directory. The next task was making the tied state triphones. One of the mechanisms provided by HHEd was employed for this purpose. It was performed by running HHEd.

The edit script tree.hed containing the instructions regarding which contexts to examine for possible clustering, was generated using the perl code provided in the HTK tutorial.
So far the set of triphones used, did cover only the training data. But here a new list of triphones was required, containing all triphones from all type of data. It was then created using the *HDMan command.*
This command generated the new list called *fulllist* that is expanded to include all the triphones needed for recognition both in the test and training data. Basically the new list was required by the AU command in the tree.hed script. This command then used the tree to synthesize all of the new previously unseen triphones in the new list.

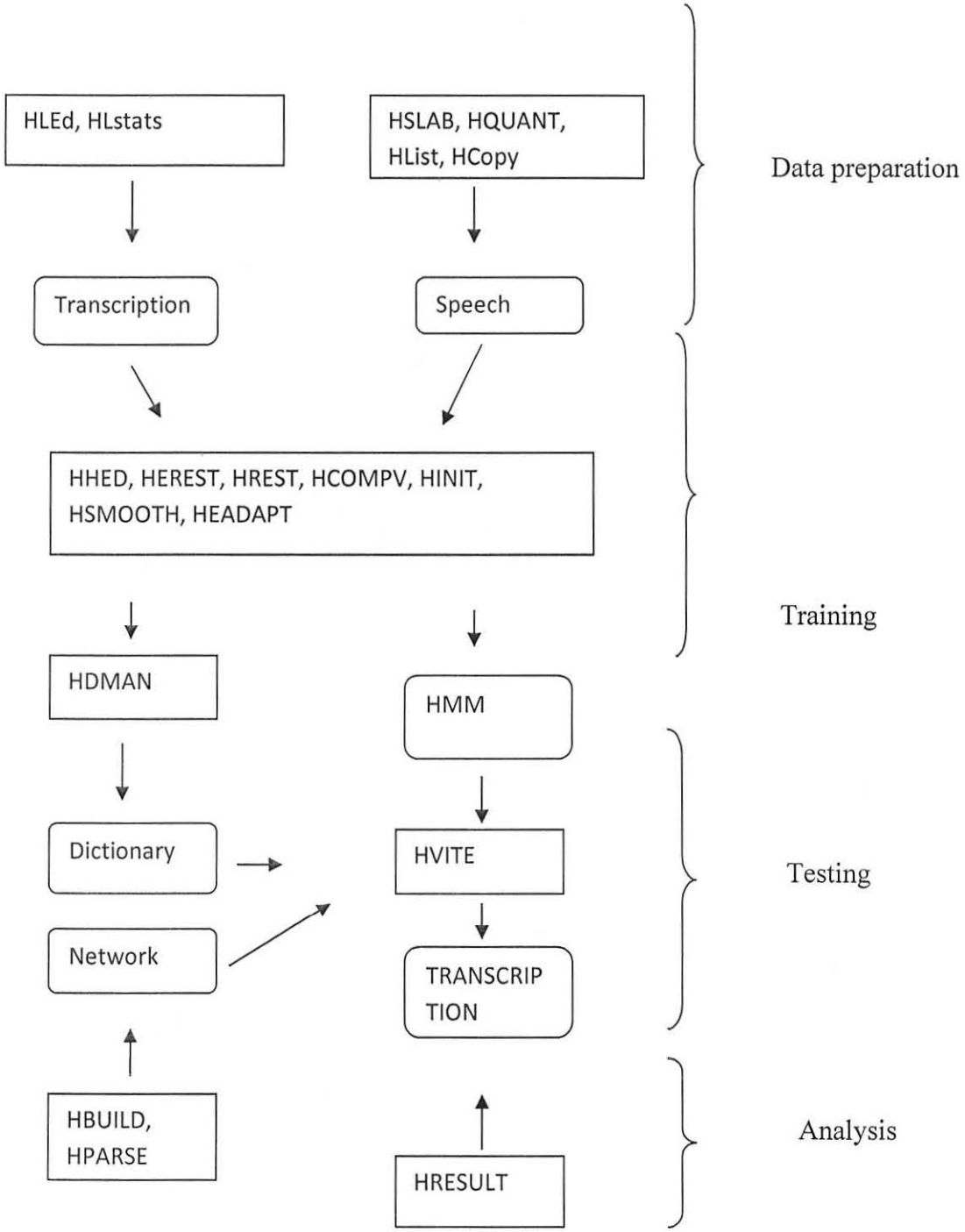

Fig.1.1. HTK processing stages

4.4. Testing and Evaluation

Evaluation and testing were the final tasks in the process of developing a speech recognizer. Therefore, in the end the performance of the developed recognizer was tested using the evaluation test set. In HTK the Viterbi algorithm is used to measure the performance of the recognizer. This algorithm is implemented by the tool HVite in HTK. The transcriptions output by the Viterbi algorithm were compared to the original word level transcription files using the HResult analysis tools. This tool uses a dynamic programming-based string alignment procedure. The analysis tool computes the percentage of words correctly recognized using the following formula: Correct $= \frac{H}{N} * 100$

Where H is the number of correct words and N denotes the total number of words in the reference transcription. In addition, the tool also performed word accuracy measure which takes into account the fact that some of the words classified as correct may be in fact insertion (I) errors. This is computed by:

Accuracy $= \frac{H-I}{N} * 100$

According to the experiment undertaken, varying results were obtained with canonical and alternate pronunciations.

The two models have different pronunciation dictionary so as to compute their respective performance level in order to deduce if it is possible to handle pronunciation variation. The result obtained from the experiment conducted is shown below.

Approaches	Sentence	Word
Canonical	%correct = 5.26 H= 13 N= 247	%correct = 44.73 Accuracy =37.21 H=1082 D =94 S = 1243 I= 182 N=2465
Knowledge based	%correct = 7.69 H= 19 N= 247	%correct = 48.75 Accuracy =41.44 H=1207 D =119 S = 1150 I= 181 N=2465
Data driven	%correct = 7.69 H=19 N=247	%correct = 50.28 Accuracy =42.9 H=1254 D =116 S = 1124 I= 1824 N=2465
Hybrid	%correct = 7.69 H=19 N=247	%correct = 50.71 Accuracy=43.57 H=1250 D= 119 S= 1096 I=176 N=2465

Table 4.1. The recognizer performance with different approaches

In the above table, H represents the number of correct words recognized, D represents the number of deletions (words that are present in the reference transcription, but are 'deleted' by the recognizer and do not occur in the recognizer's transcription), S represents the number of substitutions (words in the reference transcription that are 'substituted' by other words in the recognizer's transcription), I represents the number of insertions (words that are present in the recognizer's transcription but not in the reference), and N represents the total number of words in the reference transcription.

The test sentences used in this research were 247 consisting around 2465 words. Among these there are 200 words having variation. The variation is obtained using knowledge based, data driven approach and hybrid approach. These variants were added to the dictionary. To investigate impact of the pronunciation variation, first the sentences were tested with canonical, then after the words with pronunciation variation were added for investigation. Consequent to adding variants to the dictionary, better performance is obtained as shown in the above table. Though the same number of variants is added to both knowledge based and data driven, the obtained result is quite different. The reason is the variants added in case data driven approach actually exist while of knowledge based may or may not actually exist. The model developed using hybrid dictionary (combination of knowledge based and data driven) has relatively shown promising improvement due to the better match between the language model and the acoustic model.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1. Conclusion

As the ultimate goal of this study is handling pronunciation variation in ASR system an experiment was performed using data set prepared to achieve the objective. To arrive at appropriate results for the language under study, it was performed through integration and implementation of tools, techniques and methods for the possible outcomes of the recognizer.

This section presents conclusion drawn from the experiment done.

- Dictionary was prepared with canonical, knowledge based, data driven and hybrid approach.
- The relative performance of every model with different pronunciation dictionary was compared against one another.
- HMM modeling techniques and HTK were used for developing the recognition system.
- Phonemes were taken as recognition unit.
- Only variation within word was considered; the study didn't consider variation across words.
- Model was developed with canonical and alternate (hybrid) to check the possibility of handling variation problem. Thus for canonical an accuracy of5.26% and 44.73% for sentences and word was obtained while of the hybrid is 7.69% and 50.71% for sentences and words respectively.

Thus despite the low performance obtained, the result of the experiment is a proof of the fact that it is possible to handle pronunciation variation in Amharic speech recognizer.

5.3. Recommendation

The following points are recommended for further study to fulfill the remained gap so that to make the system fully fledged.

- It has been indicated that the recognizer demonstrated here was speaker independent but its independence is only to those people whose first language is Amharic or at least to those who can speak the language fluently. This is mainly because; the database was constructed from those native speakers. Thus, pronunciation variation due to second language should be investigated.
- The study considered only variation within words but variation that exists across words has also its own impact on ASR system performance. Thus further study is recommended for variation across words.
- As quality of the data set affects the performance of ASR, standardized data set for Amharic language is recommended as future study.
- HMM was used for modeling technique in this study, in the future other modeling techniques like artificial neural network is recommended to evaluate the performance of the model.
- In our work, we used only a bigram language model generated with HTK model, in the future higher order n-gram language models should be developed using better development tools such as SRILM.

REFERENCES

Aster Tadesse: The syllable Structure of Amharic and the Syllabification of Medial Consonant Clusters and Geminates. Addis Ababa: Addis Ababa University, (1981).

Abdulla: "Signal Processing and Acoustic Modelling of Speech Signal for Speech Recognition Systems"PhD Thesis, Information Science Department, University of Otago, New Zealand,(2002).

BayeYimam: *የአማርኛሰዋስው(YamariɲaSəwasəw)* 'Amharic Grammar book'. Revise ed. AddisAbaba: E.M.P.D.A, (2008).

Bender,L.M: Towards a Lexicostatistic Classification of Ethiopian Languages. Apaper Prepared for Colloquium on Hamito-Semitic Comparative Linguistics, 1976.

Cowley: "The Amharic Language."Languages in Ethiopia. London: Oxford University Press, (1976).

Crème,Lie: In search of better pronunciation models for speech recognition, (1996).

Crystal D:*The Cambridge Encyclopedia of Language*. Cambridge: Cambridge,University Press, (1987).

Getahun Amare: Yamariɲa Səwasəw Bək'əlal Ak'ərarəb (Modern Amharic Grammar in a Simple Approach). Addis Ababa: Commercial Printing Press, (1997)

Mulugeta Seyoum. The Syllable Structure and Syllabification in Amharic. MPhil, Thesis.Department of Linguistics, Trondheim, Norway ,(2001).

Chollet, Gerard: "Automatic Speech and Speaker Recognition: Overview, Current Issues and Perspectives." In Fundamentals of Speech Synthesis and Speech

Recognition: Basic Concepts, State of the Art and Future Challenges. pp 129-147. Edited by Eric Keller.Chichester: John Willy & Sons, (1994).

Dawkins, C: Fundamentals of Amharic.Addis Ababa Sudan Interior Mission, (1969).

Daniel Jurafsky: Speech and Language Processing: An Introduction to Natural Language Processing, computational Linguistics, and Speech Recognition, (2009).

Erwan, P: Voice, speech and gender: male-female acoustic differences and cross language. Paris: Université Paris, (2012).

Fosler-Lussier, E: *Dynamic Pronunciation Models for Automatic Speech Recognition:* Ph. D. thesis, University of Nijmegen, Netherland, (1989).

G. Chollet: Test set definition and specification. Technical Report LE2-4001-SD1.3.4, Consortium and CEC, (1994).

Helmer Strik ,Catia Cucchiarini: Modelling pronunciation variation for ASR: A survey of the literature , Department of Language and Speech, University of Nijmegen, Netherlands, (1999).

Holmes J., Holmes W: Speech Synthesis and Recognition: Second Edition NewYork, (2001).

J.R. Deller, J.G. Proakis, and J.H.L. Hansen: Discrete Time Processing of Speech Signals: Secon Ed. New York, Macmillan, (1993).

Kessens: Making a Difference on Automatic Transcription andModelling of Dutch Pronunciation Variation for Automatic Speech Recognition;Nijmegen University, Netherland,(2002).

Judith, M., Kessens, &Strik, M. W: Modeling Within-Word and Cross- Word Pronunciation Variation to Improve the Performance of Dutch ASR. Netherland: University of Nijmegen,(2002).

Kasehun Gelana: Speaker Independent, Continuous Speech Recognition for Afaan Oromo: Unpublished Msc. Thesis Addis Ababa University,(2010).

KinfeTadesse: Sub-word based Amharic speech recognizer: An experiment using Hidden Markov Model (HMM). MSc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia,(2002).

Laver, Eds., pp. 256–297. Blackwell Publ., Oxford, (1994).

Nam Soo Kim et.al., On estimating Robust probability Distribution in HMM in HMM based speech recognition , IEEE Transactions on Audio, Speech and Language processing Vol.3,No.4, July(1995).

L. R. Rabiner: A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", (1990).

Hayward, K. & R. J. Hayward. Amharic. (In) The Handbook of the International Phonetics Association, 45-50. Cambridge: Cambridge University Press.(1999).

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

Leslau, Wolf. Reference Grammar of Amharic. Wiesbaden: Harrassowitz,(1995)

L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of IEEE, (1989).

Furui, S., Digital Speech Processing, Synthesis, and Recognition Second Edition, Revised and Expanded, Marcel Dekker, Inc., New York, (2000).

.L.R.Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition , Proc.IEEE 77(2):257-286.February (1989).

Lee,Tan: Modeling Cantonese Pronunciation Variations for Large-Vocabulary Continuous Speech Recognition, the Association for Computational Linguistics

and Chinese Language Processing, (2006).

Leslau, Wolf. 1995. Reference Grammar of Amharic. Wiesbaden: Harrassowitz.

Juang and S. Furui: "Automatic recognition and understanding of spoken language – A first step towards natural human-machine communication", Proc. IEEE (to be published)

S. Furui: "Speech recognition technology in the ubiquitous/wearable computing environment", Proc.IEEE Int. Conf. Acoust., Speech, Signal Process., Istanbul, pp. 3735-3738 (2000)

Young, Steve. Large Vocabulary Continuous Speech Recognition: a Review. Cambridge University: Cambridge.(1996)

W.H. Abdulla, Auditory Based Feature Vectors for Speech Recognition Systems Electrical and Electronic Engineering Department of Auckland University, Auckland,Newzealand. ,(2002)

Lee, K-F: Context-Dependent Phonetic Hidden Markov Models for Speaker- Independent Continuous Speech Recognition. Readings in Speech Recognition, (1990).

.R.P.Lippmann, An introduction to computing with neural nets , IEEE ASSP Mag., 4(2),pp.4-22,April 1987.

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

Martha Yifiru: Automatic Amharic Speech Recognition System to Command and Control Computers. MSc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia,(2003).

Mesfin Brilie 2008. Synthetic Speech Trained - Large Vocabulary Amharic Speech Recognition System (SST-LVASR), Addis Ababa University.

Strik, H. & Cucchiarini, C. Modeling pronunciation variation for ASR: A survey of the literature. Speech Communication 29, 225-246.(1999).

MirjamWester: Pronunciation Variation Modelling for Dutch Automatic Speech Recognition; University of Nijmegen, Netherland, (2002).

Mullen, D.S: "Issues in the Morphology and Phonology of Amharic: the Lexical Generation of Pronominal Clitcs." Ph.D. Thesis. University of Ottawa, (1986)

Odelly: The Use of Context in Large Vocabulary Speech Recognition: Dissertation submitted to the University of Cambridge, London, (1995).

Amdal and Fosler-lessier "On the road to improved lexical confusability metrics,"in proc.ISCA ITRW pronunciation modeling and lexicon Adaptation (PMLA),(Ester park(co),USA),pp.53-58,(2002).

Juang: Fundamentals of Speech Recognition: Englewood Cliffs,New Jersey: Prentice Hall, Inc, (1993).

Ladefoged, P. 2001a. A Course in Phonetics. 4th Edition. Fort Worth: Harcourt, Brace, Jovanovic.

Young, The HTK Book: for HTK Version 2.0, Cambridge University Press, Cambridge,England, (1995).

S. Schotz: "Prosodic cues in human and machine estimation of female and male speaker age," in Nordic Prosody. Proc. of the IXth Conference, Lund, (2004).

Steve Young : The HTK Book (for HTK Version 3.4) Cambridge University, (2002).

Steve Young: Large Vocabulary Continuous Speech Recognition: A Review. Cambridge University, (1996).

S.K.Katti :Department of Computer Science and Engineering Sri Jayachamarajendra College of Engineering Mysore, India, (2009).

Cucchiarini : Phonetic Transcription A Methodological and Empirical Study. Ph.D. thesis, University of Nijmegen, The Netherlands. (1993)

TadesseBeyene: "The Ethiopian Writing System." Paper presented at the 12th International Conference of Ethiopian Studies, Michigan State University, (1994).

V. Mantha, R. Duncan, Y. Wu, and J. Zhao: Implementation and Analysis of speech Recognition front-ends.Mississippi State University, (2001).

Wiggers: Hidden Markov Models for Automatic Speech Recognition and their Multimodal Applictions. Delft University of Technology; the Netherlands, (2001).

Woosung Kim: Automatic Speech Recognition and Statistical Machine Translation: A Dissertation Submitted to the Johns Hopkins University in Conformity with the Requirements for the Degree of Doctor of Philosophy. Baltimore, Maryland,(2004).

Mirjam Wester. Pronunciation Variation Modeling for Dutch Automatic Speech Recognition; University of Nijmegen, Netherland, (2002).

Zegaye Seifu: Large Vocabulary Speaker Independent Amharic ASR System: Unpublished Msc. Thesis, Addis Ababa University, (2003).

Zue: Spoken Language Input: Overview. Survey of the state of the Art in Human Language Technology, (1996).

Declaration

This thesis is my original work, has not been presented for a degree in any University and all sources of material used for the thesis have been duly acknowledged.

____

Shimekit Teka

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

____

Solomon Tefera (PhD)

____

Feda Negesse (PhD)