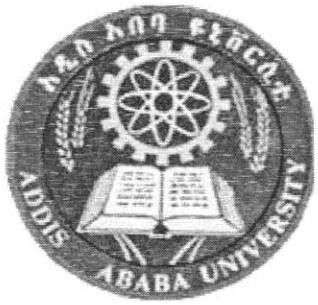


# Pronunciation Variation in Amharic Speech Recognition

## (Data-Driven Approach)

Tewedaj Alemayehu

A Thesis Submitted to

College of Humanities, language Studies, Journalism and Communications

Department of Linguistics

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science (Computational Linguistics)

Addis Ababa University

Addis Ababa, Ethiopia

June 2013

**Addis Ababa University**

**School of Graduate Studies**

This is to certify that the thesis prepared by Tewedaj Alemayehu, entitled: *Pronunciation Variation in Amharic Speech Recognition (Data-Driven Approach)* and submitted in partial fulfillment of the requirements for the Degree of Master of Science (Computational Linguistics) complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Examiner Derib Ado Signature _____ Date October 31, 2013

Examiner Million Meshesha Signature _____ Date 31/10/2013

Advisor Dr. Solomon T. Signature _____ Date 25/10/13

Advisor _____ Signature _____ Date _____

_____

Chair of Department or Graduate Program Coordinator

# ACKNOWLEDGMENT

I would like to take this opportunity to thank all who have direct or indirect relation with this thesis and study.

First of all I have to thank my creator almighty God, who enabled me to finish my study and helps me in all aspects of my life. After this I need to thank St. Mary mother of my Lord Jesus, who is always beside me. God you are good all the time and the only thing I can do is just to say thank you.

Then I forward my heartfelt thanks to my advisor Dr Solomon Teferra, who is supportive, patient and very positive who should be taken as a role-model because of his working attitude, outstanding knowledge and the value of time he gives. Beside advising and consulting me he was sharing all the important materials and resources which I used as basic input for this study. This absolutely simplified my work load. I really thank you very much Dr Solomon Teferra. I have learned a lot more than just knowledge, but a great work and study habit through your way of knowledge transfer.

Next and most importantly I am very thankful to my parents Ato Alemayehu G/Giorgies and W/ro Elfnesh W/Yohannes, who support and pray for me in all walks in my life. During my study I have also understood what matters in life is not only what to achieve, but how to achieve is the most influential point. There was a lot to be passed; therefore I have to thank my parents who are always there for me.

Finally I would like to thank Ato Onosmos Amberas and Ato Zelalem Asfaw for sharing ideas and all families, friends, relatives and Oromia Education Bureau staff for their concern and contribution during of the course study.

# Table of Contents

# List of Tables

# List of Figures

# List of Acronyms

**ASRCo**- Amharic Speech Recognition corpus

**VLCSR**-Very Large Continuous Speech Recognition

**HMM**- Hidden Markov Model

**HTK**- Hidden Markov Model Toolkit

**LM**- Language Model

**MFCC**- Mel Frequency Cepstral Coefficient

**SER**- Sentence Error Rate

**WER**- Word Error Rate

# CHAPTER ONE

# INTRODUCTION

## 1.1. General Background

### 1.1.1. Introduction to ASR Systems

For human beings, speech constitutes a very efficient means of communication, and in present days people prefer and are forced to live dependently with the help of computers. It is always natural for human beings to make life more convenient. It is this desire that contributes for innovations and discoveries of scientific and technological systems (Abdella 2010). This has induced many people to think that speech might also be a very efficient means of communication between human beings and machines. In spite of the progress that has been made, a gap still exists between the performance of human beings and machines on speech recognition.

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 (Abraham 2011). Humans usually have no difficulties in processing different pronunciation variants of the same word, since they have knowledge of pronunciation variation. However, for speech recognizers, pronunciation variation presents a problem, because, in general, speech recognizers do not explicitly take into account the different ways in which words

can be pronounced. In the beginning of ASR research, the amount of variation in pronunciation was limited by using only isolated words. Since then, the type of speech that can be processed has evolved from isolated words to spontaneous speech (Strik and Cucchinrini 2002). Especially in spontaneous speech the amount of pronunciation variation is very huge. Words are more connected to each other in spontaneous speech. As a consequence, the pronunciation of one word is influenced by that of adjacent words. Furthermore, words are usually articulated less carefully in spontaneous speech. Modeling pronunciation variation is seen as a possible way of improving the performance of ASR systems that handle spontaneous speech (Solomon 2005).

Speech is the ultimate, universal mode of human communication and it is how man should be able to interact with computers (Zegeye 2003) citing Nahm and Slater 1997). Speech interface in user's own language is in short, an ideal means of communication as being the most natural, flexible, efficient and convenient option allowing hands and eyes to be free (Nahm and Slater 1997) cited by (Zegaye 2003). It is more than half a century since computer is in use. During this period a lot of advancement has been achieved and it is believed that computers became an essential constituent in this digital world. Due to the advancements made in digital signal processing, pattern matching, classification algorithms and computer hardware technology, the dream of providing speech interface to computers has become a reality (Karl Pettey and Shneiderman 1993). Supporting this idea, cited by Adams et. al.(1999) and Martha (2003) indicated that speech technology has

advanced to "the stage where it offers great promise for human-computer interaction in a variety of applications."

Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the words that a person speaks into a microphone and convert it to written text. Speech recognition (also referred to as voice recognition) is a process by which the elements of spoken language can be recognized and analyzed, and the linguistic message it contains transposed into a meaningful form so that a machine can respond correctly to spoken words (Solomon 2005). ASR is one step in the development of an "Intelligent Machine" that can "listen" to human speech. And also Speech Recognition (SR), as described by Junqua and Haton (1996), is the "decoding of the information conveyed by a speech signal and its transcription into a set of characters." The resulting characters can be used to perform various tasks such as controlling a machine, accessing a database, or producing a written form of the input speech (Martha 2003).

The goal of an ASR system is to accurately and efficiently convert a speech signal into a text message transcription of the spoken words independent of the speaker, environment or the device used to record the speech (i.e. the microphone) (Mesfin 2008). This process begins when a speaker decides what to say and actually speaks a sentence. Speech recognition is still useful, however, because we don't need computers to actually carry on conversations with us we just need to give them commands. When you type a word or phrase, the

computer does not actually understand specific language, but it recognizes the command and software tells the computer what to do when that command is recognized. The same is true of speech recognition software. Users speak commands that are recognized by a piece of software called the *speech*-engine (Mesfin 2008). The speech engine then tells the speech application what the user said, and the application determines what to do next (Mesfin 2008). This research constitutes an attempt to indicate the impact of pronunciation variations in the development of Automatic Speech Recognition Systems which is an adequate way of decreasing Word Error Rates in order to improve the performance of ASR.

### 1.1.2. Automatic Speech Recognition Classification

As indicated by Solomon (2005), research and application effort is being exerted to develop a usable continuous speech recognition system; due to the fact that it is the continuous speech recognition system, not isolated word nor connected speech recognition systems, that enables natural man-machine oral communication. It is also essential in many applications where large populations of native users, who are neither able nor willing to speak words in isolation and clearly, interact with the recognizer.

ASR systems can be classified according to some parameters that are related to the task. Some of the parameters are (Solomon 2005):

**Vocabulary size**: speech recognition is easier when the vocabulary to recognize is smaller. For example, the task of recognizing digits (10 words) is relatively easier

when compared to tasks like transcribing broadcast news or telephone conversations that involve vocabularies of thousands of words. The number of words in the vocabulary is a constraint that makes a speech recognition system small, medium or large. As a rule of thumb, small vocabulary systems are those which have a vocabulary size in the range of 1-99 words; medium, 100-999 words; and large, 1000 words or more (Deller*et al.*1993) cited by (Abraham 2011).

Large vocabulary speech recognition systems perform much worse compared to small vocabulary systems due to different factors such as word confusion that increases with the number of words in the vocabulary. For small vocabulary recognizers, each word can be modeled. However, it is not possible to train acoustic models for thousands of words separately because we cannot have enough training speech and storage for parameters of the speech that is needed. The development of large vocabulary recognizers, therefore, requires the use of sub-word units. On the other hand, the use of sub-word units results in performance degradation since they cannot capture co-articulatory effects as words do. The search process in large vocabulary recognizers also uses pruning instead of performing a complete search. Pruning, however, increases recognition errors (Solomon 2005).

However, the vocabulary size is not a reliable measure of task complexity. The grammar constraints of the task can also affect the complexity of the system. That

is, tasks with no grammar constraints are usually more complex because all words can follow any word.

**Speaking style**: this defines whether the task is to recognize isolated words or continuous speech. In isolated word (e.g., digit recognition) or connected word (e.g., sequence of digits that form a credit card number) recognition, the words are surrounded by pauses (silence). A speech recognizer can be developed to recognize only read speech or to allow the user speak spontaneously. The latter is more difficult to build than the former due to the fact that spontaneous speech is characterized by false starts, incomplete sentences, unlimited vocabulary and reduced pronunciation quality (Solomon 2005). This type of recognition is easier than continuous speech recognition because, in the latter, the word boundaries are not so evident. In addition, the level of difficulty varies among the continuous speech recognition due to the type of interaction. That is, recognizing speech from human-human interactions (recognition of conversational telephone speech, broadcast news) is more difficult than human-machine interactions (dictation software). In read speech or when humans interact with machines, the produced speech is simplified (slow speaking rate and well articulated) so that it is easy to understand it.

**Speaker mode**: the recognition system can be used by a specific speaker (speaker dependent) or by any speaker (speaker independent). Despite the fact that speaker dependent systems require to be trained on the user, they generally

achieve better recognition results (there is no much variability caused by the different speakers). Given that speaker independent systems are more appealing than speaker dependent ones (no training required for the user), some speaker-independent ASR systems are performing some type of adaptation to the individual user's voice to improve their recognition performance.

Lee (1989), has pointed out that speaker dependent systems are useful for some applications. However, they have many problems. First the training is inconvenient and time taking to the user. Second a large amount of processing is required. Third, while certain applications may involve only multiple speakers, certain other applications, like telephone directory assistance, may not tolerate the training delay; considerable additional storage is needed if each speaker's parameters are to be stored separately; and as a speaker's voice may change over time due to stress, fatigue, sickness, or variations in microphone positioning, the system is in trouble (Zegaye 2003).

**Channel type**: the characteristics of the channel can affect the speech signal. It may range from telephone channels (with a bandwidth about 3.4 kHz) to wireless channels with fading and a sophisticated voice. Speech recognizers may require the speech to be clean from environmental noises, acoustic distortions, microphones and transmission channel distortions or they may ideally handle any of these problems (Solomon 2005). While current speech recognizers give acceptable performance in carefully controlled environments, their performance

degrades rapidly when they are applied in noisy environments. This noise can take the form of speech from other speakers; equipment sounds, air conditioners or others. The noise might also be created by the speaker himself in a form of lip smacks, coughs or sneezes (Abraham 2011).

**Transducer type**: defines the type of device used to record the speech. The recording may range from high-quality microphones to telephones (landline) to cell phones to array microphones (used in applications that track the speaker location).

One or more of these constraints can be put on a speech recognizer. Ideally, a speech recognition system should be free from any constraint. Therefore, for speech recognition systems to be beneficial and universally applicable, the constraints should be as few as possible. They should also be able to adapt themselves and learn new lexical, syntactic, semantic, and pragmatic information, just as humans do. When placed in this perspective, the field of speech recognition is seen to be in its early stage of infancy (Deller, Proakis and Hansen 1993) cited by (Solomon 2005).

## 1.2. The Amharic Language

Amharic is the working language of the Federal Government of Ethiopia. It is a Semitic language that has the greatest number of speakers after Arabic. The data that the researcher found from the Federal Democratic Republic of Ethiopia

Population Census Commission indicates that the population distribution for Amhara Regional State is 17,214,056 on the year 2007 (PCC 2007). This doesn't mean that these are the only people who speak Amharic language. It is believed that in Amhara regional state there are some people who speak other language as a mother tongue but use Amharic as their second language. In addition to this most people speak Amharic as their mother tongue or second language outside of the Amhara regional state.

Amharic has five dialectical variations spoken named as: Addis Ababa, Gojam, Gonder, Wollo and Menz (Zelalem 2007). The Amharic writing system uses multitudes of ways to denote compound words and there is no agreed upon spelling standard for compounds. As a result of this and of the size of the country leading to vast dialectal dispersion, lexical variation and homophony is very common. 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. For various reasons, words in Amharic are pronounced differently and varied from one speaker to another and from one situation to another. This is one of the major factors that degrade the performance of Amharic ASR systems. Also pronunciation variation modeling has been studied in the field of speech Synthesis and Recognition to improve performance of the corresponding speech systems for other languages like English (Abraham 2011).

## 1.3. Statement of the Problem

It has been more than five decades since Research in ASR started in Europe and USA (Kathy, Gorang, Ralph, Cecilia, Brenton, Luis, 2007). In Ethiopia ASR Research in local languages was started when Solomon (2001) showed the possibility of developing Automatic Speech Recognition for Amharic using Isolated Consonant-Vowel Syllable Recognition System. This is followed by other research works aiming for Amharic Automatic Speech Recognition systems and other local languages such as Afan Oromo, Sidama, Tigrigna etc, which definitely shows the improved interest of researcher's in the area of Speech Recognition.

However, since Amharic is one of the under resourced language, this is limitation of pre-stored resources for the system such as speech corpus, pronunciation dictionary, language model, acoustic model application of ASR for Amharic language that needs much more research work to reach to an efficient and accurate Automatic Speech Recognizer for the Language.

In any language, speech is the most natural means of communication between human beings. One of the first skills we learn to use from our babyhood and which can be done without any tools or any explicit education is speech. It is obvious that before the invention of writing, speech was the only way of passing knowledge. Even in the modern world, despite all our novel ways of communicating, the *speak-listen* style is opted most. It is also the most important way of communication. When speaking with somebody, one does not have to

focus on the audience; he/she can look in a different direction or even perform some other task while communicating. So it is only logical that machine interface designers in their quest for a natural man-machine interface have turned to automatic speech recognition and speech production as one of the most promising interfaces (Wiggers 2001 cited in Zegaye 2003).

When dialogues between humans and computers are more natural, the ASR must handle more conversational speech. Conversational speech is harder for ASR systems to recognize correctly, because of increased co-articulation and pronunciation variability, as well as less predictable language usage. Weintraub et al (1996) showed that a spontaneous speaking style is harder to recognize; when the same exact word sequences were recorded in a truly spontaneous, acted spontaneous, and read style, the ASR system performed much worse on spontaneous speech compared with the other two styles. Some of the pronunciation variation is caused by speaking style (dialects, non-native mother tongue, etc.), and may be better handled by careful design of the pronunciation dictionary, i.e. pronunciation modeling.

The most common way of dealing with pronunciation variation is to put several pronunciation alternatives in the ASR lexicon. These pronunciations are also often used to re-transcribe the speech corpus before a retraining of the acoustic models. Using the lexicon to capture speaker variation makes it possible to model

several speakers simultaneously, thus using the same lexicon and the same acoustic models for all speakers (Muhirwe 2005).

High quality recognizers always include a language model, which is sometimes incorporated in pronunciation modeling techniques. For large vocabulary speech recognition, a well-designed language model may decrease the negative impact of a mismatch between the speaker and the acoustic models and explicit pronunciation modeling may be less important. If the speaking style we try to model has special language model characteristics, e.g. the hesitations and restarts of spontaneous speech, they may be incorporated directly into the language model. Even if it has been done lot; Automatic Speech Recognition systems have a limitation of recognizing possible pronunciation variation of different peoples. Current Automatic Speech Recognizers do not know the same word with different reading pronunciation, which increases the word error rate of the system (Ingunn and Eric 2003).

The development of a large vocabulary speaker independent recognition system requires the availability of an appropriate pronunciation dictionary that encompasses a large number of words with their pronunciation. The pronunciation dictionary, which is the lexical model, is one of the most important blocks in the development of large vocabulary speaker independent recognition systems. A pronunciation dictionary is a machine-readable transcription of words in terms of sub-word units. It specifies the finite set of words that may be output

by the speech recognizer and gives, at least, one pronunciation for each. A pronunciation dictionary can be classified as a canonical or alternative dictionary on the basis of the pronunciations it includes (Solomon 2005).

The objective of ASR is to derive the correct string of spoken words from an acoustic signal. However, pronunciation variation makes it more difficult to achieve this objective, as the variation can result in recognition errors. The goal of pronunciation modeling is to minimize the recognition errors due to pronunciation variation and thus improve the performance of the ASR system. The recognition errors can be a direct result of variants that are pronounced but not included in the lexicon (Ingunn and Eric 2003).

Therefore, this research aimed at showing the effect of pronunciation variations by using directions in finding pronunciation variations, of a word that exist in the research corpus of Solomon (2005) using the d*ata-driven methods.* The problem is that the variations based on a given database may give a result too specific for that database. One of the advantages is that we may compute probabilities for the variants, as opposed to the knowledge-based methods and later improves the performance of the system.

## 1.4. Research Questions

- Does pronunciation variation have an impact on the performance of Amharic Speech Recognition?

- ❖ Does adding variants to the pronunciation dictionary improve the performance of Amharic Speech Recognition using data-driven approach?

## 1.5. Objective of the SStudy

### 1.5.1. General Objectives

The main objective of this research is to ***explore*** and show the effect of pronunciation variation including investigation of the contribution of pronunciation dictionary for Amharic Speech Recognition System using data-driven pronunciation variation modeling.

### 1.5.2. Specific Objectives

The specific objectives of this research are:

- To show possible pronunciation variations in a Speech Corpus.
- To show the effect of pronunciation variation for the efficiency and performance of ASR in Amharic.
- To integrate multiple pronunciations into Amharic lexicon using data-driven approach.
- To integrate acoustic evidence for Amharic pronunciation lexicon.
- To test and analyze the performance of speech recognizer model of the Amharic ASR model.

## 1.5. Scope and Limitation of the Study

The scope of this study is limited to show the effect of pronunciation variation in the performance Amharic Speech Recognition System due to inter-speaker variation so that future researches will consider the pronunciation variation to increase the performance of large vocabulary, continuous, speaker independent Amharic speech recognition system. In this research the data-driven approach is used to model the pronunciation variations for the experiment in the alternative dictionary.

## 1.6. Methodology

In order to achieve the goal of this research the researcher used the following methodologies.

### 1.6.1. Data Collection

A speech corpus is one of the fundamental requirements for any speech recognition research. The speech corpus is a collection of speech recordings which is accessible in computer readable form, and which has an annotation and documentation sufficient to allow re-use of data. The speech corpus is designed according to best-practice guidelines established for other languages. Standard speech corpora consist of a training set and evaluation test sets. The training set is

intended to collect speech data for training the recognizer and the evaluation test set is for the purpose of final evaluation of the recognizer (Abraham, 2011).

Solomon (2005) citing Schiel et.al(2003) mentions the following styles of speech that differentiate a speech corpus from others:

- Reading Speech
- Answering Speech
- Command/Control Speech
- Descriptive Speech
- Non-prompted Speech
- Spontaneous Speech
- Neutral vs. Emotional Speech

The preparation of any type of speech corpus is normally a research/project by itself, which includes all the knowledge, finance and time aspects of a research.

As cited by Solomon (2005), Schiel et.al(2003) pointed out that most speech corpora contain read speech, either for practical reasons because annotating non-read speech is more difficult, or simply, because the intended application or investigation requires read speech. Due to this read corpus is the relevant format for this research.

The preparation of a read speech corpus usually involves the following steps:

1. Suitable training and test sentences are selected from a database of sentences to be used as prompts for the speaker;
2. The selected text sentences are read aloud by a number of chosen speakers and recorded;
3. The recorded speech is preprocessed, i.e., it is transcribed and segmented;
4. Proper documentation.

Here, the researcher used an Amharic text corpus prepared by Solomon (2005). The Amharic Speech Corpus of Solomon (2005) has been designed according to best-practice guidelines established for other languages. Standard speech corpora, such as the Wall Street Journal (Frasen 1994), consist of a training set, a speaker adaptation set, development test sets (for 5,000 vocabulary and 20, 000 vocabulary), and evaluation test sets (for 5, 000 vocabulary and 20, 000). The Amharic corpus has, therefore, been made to contain the same components.

### 1.6.2. Data Sampling

In this research we have selected 10 speakers for the training and 5 speakers for the testing, both female and male individuals by considering large vocabulary continuous speaker independent ASRS. Speakers were selected randomly from the speech corpus based on the researchers hearing ability. There was no specific or special reason for the selection of the individuals. As the aim of this research is to show pronunciation variations, the focus was only how the words were

pronounced in different individuals. Therefore, we have selected and used total of 471 sentences with 2887 words and 3655 pronunciations for the training and 90 sentences with 847 words and 1022 pronunciations were used for the test.

### 1.6.3. Modeling

The researcher used the Hidden Markov Model (HMM) which is widely known and used for speech recognition systems and developed an alternative pronunciation for the data driven approach which was not considered in previous researches of Automatic speech recognition researches for Amharic language.

### 1.6.4. Performance Measure

Performance measure was done after all necessary which are data preparation, training were accomplished. HVite is a tool that was used to recognize the pronunciation dictionary, language models and an output of a transcription file against which the recognizer's performance was analyzed in HTK. Both the canonical and the alternative (multiple) pronunciation dictionaries performance was measured in the testing steps.

The recognition result of the canonical and alternative dictionary of the Amharic speech recognition system pronunciation variation was evaluated by the command HResult, which finally gives the word accuracy, word correctness, sentence accuracy and sentence correctness in percents manner.

## 1.7. Expected Benefits of the Study

The significance and benefit of this research is to contribute for the Amharic language in the natural language processing system. Thus the research has prominent significance to help Automatic Speech Recognition System improvement by exploring the effect of pronunciation variations and showing a model of data-driven pronunciation dictionary using the Corpus of Solomon (2005).

The researcher also tested the performance of the recognizer with the pronunciation dictionary to see the result of word error rate to show the contribution of dealing with variations of pronunciations in automatic speech recognition efficiency increase. Therefore, the benefit is that after words researchers in the area of automatic speech recognition will consider possible pronunciation variations in the development Amharic ASR which will improve the performance of the system.

## 1.8. Organization of the Thesis

This thesis is organized into six chapters, chapter one discuses about introduction to ASR systems, classification of Automatic Speech Recognitions, overview of Amharic language, data sampling, modeling, performance measure and expected benefit of the study.

Chapter two presents about basic concepts of Automatic Speech Recognition especially focusing on Amharic Speech Recognition in point of view of the impact of pronunciation variations and shows what has been done by researchers or related works on Amharic Automatic Speech Recognitions.

Chapter three discusses about the Amharic language, including linguistic features of the Amharic language, the writing system/transcription, phonetics, phonology, morphology, syntax, semantics and pragmatics aspects.

Chapter four presents about issues and major components of Automatic Speech Recognition, errors made by speech recognizers, HMMS, and approaches to the development ASR, types of HMMs, pronunciation variation, pronunciation variation modeling for ASR and issues in pronunciation variation modeling.

Chapter five shows effects and results of pronunciation variation in Amharic Speech Recognition with experimental Analysis which shows the analysis and experimental results of the pronunciation variations effects in the development of Automatic Speech Recognition.

Chapter six concludes what has been found in the research and gives recommendations for future works.

# CHAPTER TWO

# LITERATURE REVIEW

## 2.1. Pronunciation in Natural Language

Speech is a process used to communicate from a speaker to a listener (Strik and Cucchiarini 1999). Pronunciation relates to speech, and humans have an intuitive feel for pronunciation. 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, which deals with the different ways of speaking a given word. 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.

Communication using speech is inherently natural, with this ability of communication unconsciously acquired in a step-by-step manner throughout life. Speaking and comprehending speech can be viewed as speech chain, a kind of "brain-to-brain" linking. The grammar relates sounds and meanings, and contains the units and rules of the language that make speech production and comprehension possible. However, other psychological processes are used to produce and understand utterances. There are mechanisms that enable us to break the continuous stream of speech sounds into linguistic units (Inguun and Eric 2003) such as phonemes, syllables, and words in order to comprehend, and to

compose sounds into words in order to produce meaningful speech. Other mechanisms determine how we pull words from the mental lexicon, and still others explain how we construct a phrase structure representation of the words we retrieve.

We usually have no difficulty in understanding or producing sentences in our language. We do it without effort or conscious awareness of the processes involved. However, we have all had the experience of making a speech error, of having a word on the "tip of our tongue," or of misunderstanding a perfectly grammatical sentences. The human brain is able not only to acquire and store the mental lexicon and grammar, but also to access that linguistic storehouse to speak and understand language in real time (Warner 2001).

## 2.1. Pronunciation Variation and Automatic Speech Recognition

Pronunciation variation refers to the fact that words can be pronounced in many different ways (Ingunn and Eric 2003). 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 or minor difficulties in processing different pronunciation variants of the same word, since they have knowledge of pronunciation variation. However, for speech recognizers, pronunciation variation forms a problem, because, in general, speech recognizers do not explicitly take into account the different ways in which words can be pronounced. In the beginning of ASR research, the amount of

variation in pronunciation was limited by using only isolated words. Since then, the type of speech that can be processed has evolved from isolated words to spontaneous speech. Especially in spontaneous speech the amount of pronunciation variation is very large. Words are more connected to each other in spontaneous speech. As a consequence, the pronunciation of one word is influenced by that of adjacent words. Furthermore, words are usually articulated less carefully in spontaneous speech (Kessen 2002).

Many factors have an impact on the pronunciation of a given word, (Ingunn 2002) for example, the position of the word within the utterance, dialect, age, and gender of the speaker. Pronunciation variations are common sources of recognition errors in real-world application of Automatic Speech Recognition, so that specific techniques must be developed to handle them (Ingunn 2002). When dialogues between humans and computers are more natural, the ASR must handle more conversational speech. Conversational speech is harder for ASR systems to recognize correctly, because of *increased co-articulation and pronunciation variability*, as well as *less predictable language usage.*

Early automatic speech recognition (ASR) systems only considered restricted speaking styles, i.e. careful articulation of isolated or connected words (Markus 2003). The increased modeling capacities of current ASR systems also manage the looser articulation of continuous speech. Making speech technology based applications more widespread has several consequences for the demands on ASR

systems (Kessen 2002). New and larger vocabularies are needed when ASR systems are used in new domains. When the vocabulary is increased it is no longer feasible to select the pronunciation variants by hand. Pronunciations from various sources will often be combined and a decision must be taken whether to include one or more variants per word. The less restricted grammar of a large vocabulary speech recognition system will give more confusability, and more care must be taken in the selection of pronunciation. Ideally, speech recognizers should handle these diverse speaking styles, (e.g. spontaneous speech, hyper-articulated speech, accents, dialects, and speech from users with different mother tongues).

All ASR systems must handle variations. The same word spoken several times by the same speaker will vary both in length and acoustical content. For speaker independent speech recognition the voice quality and characteristics will vary even more. As indicated by Ingunn (2002:3) the variation in speech input to a speech technology based service may be divided into three groups:

1. Pronunciation variation
2. Grammar and vocabulary variation
3. Channel and noise variation

Pronunciation, grammar and vocabulary variation will be speaker dependent where as channel and noise variation will depend on the environment. Trying to make ASR handle both speaker and environment variation is crucial in automatic

speech recognition which considers variation beyond the inter-speaker and intra-speaker acoustic differences even for read speech modeling.

Different speakers using different speaking styles will use different pronunciations. Spontaneous speech and different dialects or accents are examples of speaking styles with pronunciations that differ from the canonical ones often found in pronunciation dictionaries. The population of most countries becomes more and more multinational, and non-native users will increase the observed variability in pronunciation even more. There will also be differences between expert and novice users of a speech based service (e.g. Fast versus over-articulated speech). User-friendly systems should be able to recognize the pronunciations judged appropriate by the user. The use should be spared surprising system actions when using "non-surprising" speech. This will help the user to keep a consistent mental model of the system, which is of great importance for a well-designed dialogue system. If the recognizer makes errors when the user is using rare words or pronunciations, this will be understandable for the user. To make dialogue systems using speech recognition more user-friendly, robustness to common pronunciation variation is needed (Ingunn 2002).

The task that the speech based application intended gives requirements for the recognizer's vocabulary and grammar, and presents another source of variation. The vocabulary and grammar preferred by the user may vary dependent on e.g. non-nativeness, dialect and sociolect, as well as differences between expert and

novice users. One cannot assume that the users of a speech technology application will stick to a well-defined grammar. Users may be unwilling to normalize both pronunciation and grammar. They may also be unaware of their own peculiarities. Many perceive their own speaking style as normalized but all these "normalized" variants differ (Ingunn 2002).

The third main variable that also needs to be addressed in speech recognition-based applications is non-speech variation, e.g. noise and channel variation. Both pronunciation and grammar variation are dependent on the user and therefore quite different from this last type of variation that is dependent on the environment. To control how we model the observed variation, we should treat the environment and speaker variation separately (Ingunn 2002). But here the researcher's main focus is to show the influence and/or effect of pronunciation variation in Automatic Speech Recognition which will decrease the word error rate that implies the increase of systems efficiency for Amharic language which has a consonant-vowel syllable feature.

Speech recognition transcribes natural speech while speech understanding extracts the meaning of the speech. Recognizing and understanding a spoken sentence is obviously a knowledge-intensive process which must take into account and process different aspects of the speech communication process, including acoustics, phonetics, syntax, semantics and pragmatics amongst other things.

However, approaches to automatic speech recognition are limited in their ability to handle all these aspects of speech communication process (Solomon 2005).

Automatic speech recognition (ASR) can also be defined as the independent, computer-driven transcription of spoken language into readable text in real time (Stuckless 1994). In a nutshell, ASR is technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to written text. Having a machine to understand fluently spoken speech has driven speech research for more than 50 years. Although ASR technology is not yet at the point where machines understand all speech, in any acoustic environment, or by any person, it is used on a day-to-day basis in a number of applications and services. The ultimate goal of ASR research is to allow a computer to recognize in real-time, with 100% accuracy, all words that are intelligibly spoken by any person, independent of vocabulary size, noise, speaker characteristics or accent (Kessen, Wester and Strik 1999).

Even though speech recognition uses some of the same fundamental technology as voice recognition, it is different because it does not try to identify individuals. Rather it tries to recognize what individuals say. It is the difference between knowing *who* is speaking and *what* is said (Ibrahim 2011). Speech recognition is also different from voice recognition, though sometimes people may use the terms interchangeably. In a technical sense, voice recognition is strictly about trying to recognize individual voices, not what the speaker said. It is a form of

biometrics, the process of identifying a specific individual, often used for security applications.

One of the major difficulties in speech recognition systems is the variability in speech data, due, among other reasons to alternate pronunciation of words, even within the same speaker (Wooters and Stolke 1996). The lexicon is usually composed of a set of words and a single pronunciation for each of them. This pronunciation is considered to be the standard or correct one, but it usually doesn't have to do very much with the actual pronunciation of the word in a real environment (Westendorf and Jelitto 1996) cited by (Abraham 2011).

Research in Automatic Speech Recognition for Amharic was first conducted by Solomon (2001) by showing the possibility of development of ASR. He investigated that the Amharic language possesses unique features including the CV syllables which are basic units of both writing and speaking. He demonstrated a HMM-based Amharic CV syllable recognition system using the HTK toolkit. The wave files of input utterances were converted into parameterized form using an MFCC. A three state left-to-right HMMs were trained for each of the 41 selected CV syllables and a silence (sil). The models were developed for both speaker dependent and speaker independent prototypes. Performance evaluation tests were made using the test data set prepared for the research purpose. The test results were an average recognition accuracy of 87.68% obtained for the speaker dependent systems. An average recognition accuracy of 72.75% was obtained

when the test data set is from those people participated in the training. A result of 49.21% was also obtained when the test data set was from people who did not participate in the training. Given the time constraint and the result obtained, it can be concluded that the development of Amharic CV syllable recognition is feasible using the current HMM setting.

Kinfe (2002) developed sub-word based isolated word recognition systems for Amharic using HTK. The sub-word units used in the experiment are phones, triphones and CV-syllables. It considered 20 phones (out of 39) and 104 CV syllables, which are formed using the selected phones. Speech data of 170 words, which are composed of the selected sub-word units, have been recorded by 20 speakers (speech of 15 speakers for training and the remaining for testing). Speaker dependent phone-based and triphone-based systems had an average recognition accuracy of 83.07% and 78% respectively. Phone-based and triphone-based speaker independent systems had an average recognition accuracy of 72% and 68.4% respectively. In addition, a comparison of the different sub-word units revealed that the use of CV syllables has led to relatively poor performance.

Zegaye (2003) investigated the possibility of developing large vocabulary, speaker independent and continuous speech recognizer for Amharic language. The recognizer was developed using both phone-based and tri-phone based recognizers using HTK. For the training and testing of its recognizers, it used the speech data, which consists of 8000 sentences read by 80 people, recorded by

Solomon *et al.* (2005). In order to support the acoustic models, a bigram language model was constructed. In addition, pronunciation dictionary was prepared and used as an input. Since the recognizer was meant to recognize large vocabulary and continuous speech, phonemes were opted as the basic unit of recognition. The best recognizer was a tri-phone based recognizer which has 76.2% word recognition accuracy.

Martha (2003) explored the possibility of developing an Amharic speech input interface to command and control Microsoft Word. It required a speech recognizer and a communication interface between the recognizer and the application. 50 command words were selected from different menus (File, View, Insert, Tools, Table, Window, and Help), translated to Amharic and used to develop the prototype system. To develop and test the required Amharic speech recognition system, speech data were recorded from 26 people (10 female and 16 male) in the age range of 20 to 35. 76.9% of the recorded data were used to train the recognizers and the remaining data were used for testing the performance of recognizers. To test the performance of the system, 18 randomly selected command words were given to 6 people (3 command words for each) and these people were asked to command Microsoft Word orally. The system performed 16 commands accurately and only two command words were wrongly recognized and thus Microsoft Word performed wrong actions.

Solomon (2005) developed Automatic Speech Recognition for Amharic and also developed an Amharic speech corpus that can be used for various kinds of investigations in the development of ASR for Amharic. It explored various possibilities for developing a Large Vocabulary Speaker Independent Continuous Speech Recognition System for Amharic. The work assumed that, due to their highly regular consonant vowel (CV) structure, Amharic syllables lend themselves to be used as a basic recognition unit. Indeed, it has been able to show that syllable models can be used as a competitive alternative to the standard architecture that is based on triphone models. The acoustic model of the syllable-based recognizer requires 15MB memory. Together with the language model and use of speaker adaptation, it had a word recognition accuracy of 90.43% on the 5,000 words evaluation test set at a speed of 2.4 minutes per sentence. While the acoustic model of the triphone-based recognizer requires 38MB memory and had a word recognition accuracy of 91.31% on the same test set at a speed of 3.8 minutes per sentence. Although this is the state-of-the-art recognition performance in Amharic, it still sees the room for improvement because the word recognition accuracy of ASR in the technologically favored languages is approaching to 100%.

Hussein and Gamback (2005) developed an Amharic speaker independent continuous speech recognizer based on an HMM/ANN hybrid approach. The model was constructed at a context dependent phone level with the help of the CSLU Toolkit. It used part of the data (5000 sentences) recorded by Solomon *et al.*

# ABSTRACT

The main purpose of this research was to show the effects of pronunciation variation in large vocabulary continuous speaker independent Amharic speech recognition system performance evaluation.

Pronunciation variation refers to the fact that words can be pronounced in many different ways. Humans usually have no difficulties in processing different pronunciation variants of the same word, since they have knowledge of pronunciation variation. However, for speech recognizers, pronunciation variation presents a problem, because, in general, speech recognizers do not explicitly take into account the different ways in which words can be pronounced. The goal of an ASR system is to accurately and efficiently convert a speech signal into a text message transcription of the spoken words independent of the speaker, environment or the device used to record the speech. For various reasons, words in Amharic are pronounced differently and varied from one speaker to another and from one situation to another.

We started our experimental analysis by training the canonical dictionary, which is the baseline in ASR, and then we have found a result which 60.54% sentence accuracy and 73.23% word accuracy. Then we have transcribed the alternative pronunciations manually using the direct data-driven approach and 62.68% Sentence accuracy and 74.57% word accuracy was found for the multiple-pronunciation dictionary. Finally we conclude that pronunciation variations show performance evaluation difference in Amharic speech recognition systems.

(2005). The recognizers were tested with a total of 20 sentences read by 10 speakers (2 sentences each) who also read the training data. When the same recognizer was tested for another ten speakers who were not involved in the training, the recognition rate degraded. The best result obtained with this test data was 74.28% word recognition accuracy. There was 4.28% and 4.37 word and sentence recognition accuracy reduction respectively.

Solomon (2008) developed Multiple Pronunciation Model for Amharic Speech Recognition System. The research had tried to show the pattern variations of sound units in Amharic language for multiple pronunciation models. These are variation of sound units at lexical level due to dialects. After that an attempt to build a pronunciation dictionary for Automatic Speech Recognition (ASR) were considered. The read speech were tried to cover pronunciation variations of all dialects of the language. The study also used the corpus to develop phone based multiple pronunciation ASRS for Amharic which had word and sentence recognition accuracy of 52 %. And finally he showed the potential of developing Amharic ASRS using HTK that applies the HMM framework was demonstrated by the study. It was also proved that it is possible to use multiple pronunciation dictionary as a basic unit of phone based speech recognition for Amharic. For the study's experiment it was found that the optimal HMM topology for Amharic phone based model with five emitting states and twelve Gaussian Mixtures. Based on the findings of the study's experiments, he concluded that multiple

pronunciation model can improve the performance of the phone based recognizer.

Abraham (2011) developed Enhanced Amharic Speech Recognition System. The first method was a grapheme based canonical pronunciation dictionary that contains a single pronunciation for each word in the lexicon. The second method was a grapheme based multiple pronunciation dictionary that contains alternate pronunciations for some of the words in the lexicon. The pronunciation variants in the second method were generated by applying Amharic linguistic literature and dictionary to the words in the canonical pronunciation dictionary. All the words in these two methods were transcribed using transliteration schemes. The third method was a grapheme based multiple pronunciation dictionary where the transcriptions of words that were acquired from acoustic evidence for all words found in the second method grapheme based multiple pronunciation dictionary. Using the second and third methods led to a larger improvement in SER compared to the benchmark first method. The SER rates measured for the first method were 39%, 41%, 42% and 44% for speaker1, speaker2, speaker3 and speaker4 respectively. The SER rates measured for the second method are 31%, 33%, 35% and 38% for speaker1, speaker2, speaker3 and speaker4 respectively. Compared to the first method, a statistically significant decrement of 8%, 8%, 7% and 6% SER is measured in the second method for speaker1, speaker2, speaker3 and speaker4 respectively. Using the third method for only one of the four speakers has led to a 6% SER which was a further decrement of 25% SER compared to the second

method. Using the acoustic evidence transcription of this speaker to the other three speakers led to 12%, 17% and 19% SER for speaker2, speaker3 and speaker4 respectively. Compared to the second method, a statistically significant decrement of 21%, 18% and 19% SER was measured in the third method for speaker2, speaker3 and speaker4 respectively. Finally he described how the performance of a speaker dependent continuous Amharic speech recognizer was enhanced by modeling pronunciation variation.

Automatic Speech Recognition for other local languages of Ethiopia was also developed by the following persons; Abdella (2010) developed Speaker Dependent Speech Recognition for Sidaama Language. Habtamu (2010) also developed Speaker Dependent Speech Recognition for Wolayita Language. Kassahun (2010) developed a Continuous, Speaker Independent Speech Recognizer for Afaan Oromo. Teferi (2010) developed Speech Recognition for Afan Oromo Using Hibrid Hidden Markov Models and Artificial Neural Networks. This shows that there exists great initiation of researchers towards Speech Recognition. But still it needs much more effort because it was conducted only for some languages of the country when we compare it with the number of the languages in Ethiopia.

This research is original work of the researcher and it will have continuation considering the data size, possible pronunciation variations of the Amharic language etc.

## 2.3. Issues in Automatic Speech Recognition Systems

### 2.3.1. Theory of Speech Recognition Systems

The general problem of automatic transcription of speech by any speaker in any environment is still far from solved. But recent years have seen ASR technology mature to the point where it is viable in certain limited domains. One major application area is in human-computer interaction. While many tasks are better solved with visual or pointing interfaces, speech has the potential to be a better interface than the keyboard for tasks where full natural language communication is useful, or for which keyboards are not appropriate. This includes hands- busy or eyes-busy applications, such as where the user has objects to manipulate or equipment to control (Abraham 2011).

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).

In some applications, a multimodal interface combining speech and pointing can be more efficient than a graphical user interface without speech. The two words

were so much alike. True spoken language understanding is a difficult task and it is remarkable that humans do as well at it as we do. The goal of automatic speech recognition (ASR) research is to address this problem computationally by building systems that map from an acoustic signal to a string of words. Automatic speech understanding (ASU) extends this goal to producing some sort of understanding of the sentence, rather than just the words (Solomon 2005).

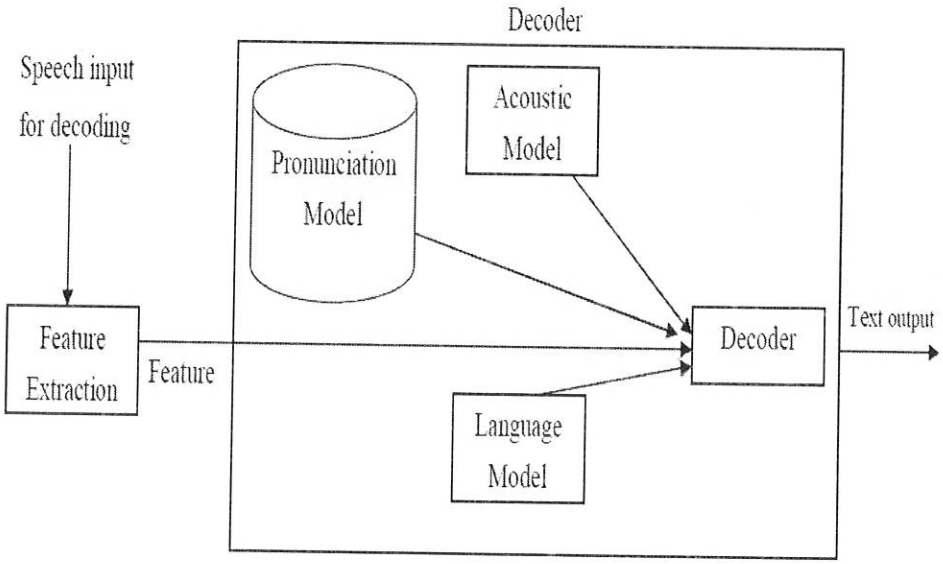


Figure 2.1. General Architecture of an ASR system (Adapted from Abraham (2011)).

### 2.3.2. Fundamentals of Speech Recognition

We have defined speech recognition in different terms and again, speech recognition is a multileveled pattern recognition task, in which acoustical signals are examined and structured into a hierarchy of sub-word units (e.g., phonemes), words, phrases, and sentences. Each level may provide additional

temporal constraints, e.g., known word pronunciations or legal word sequences, which can compensate for errors or uncertainties at lower levels. This hierarchy of constraints can best be exploited by combining decisions probabilistically at all lower levels, and making discrete decisions only at the highest level. The elements are as follows:

**Raw Speech**. Speech is typically sampled at a high frequency, e.g., 16 KHz over a microphone or 8 KHz over a telephone. This yields a sequence of amplitude values over time.

**Signal Analysis.** Raw speech should be initially transformed and compressed, in order to simplify subsequent processing. Many signal analysis techniques are available which can extract useful features and compress the data by a factor of ten without losing any important information. The most popular signal analysis techniques are the following (Jurafsky 2009).

- Fourier analysis (FFT) yields discrete frequencies over time, which can be interpreted visually. Frequencies are often distributed using a *Mel* scale, which is linear in the low range but logarithmic in the high range, corresponding to physiological characteristics of the human ear.

- Perceptual Linear Prediction (PLP) is also physiologically motivated, but yields coefficients that cannot be interpreted visually.

- Linear Predictive Coding (LPC) yields coefficients of a linear equation that approximate the recent history of the raw speech values.

- Cepstral analysis calculates the inverse Fourier transform of the logarithm of the power spectrum of the signal.

**Speech frames.** The result of signal analysis is a sequence of *speech frames*, typically at 10 m/sec intervals, with about 16 coefficients per frame. These frames may be augmented by their own first and/or second derivatives, providing explicit information about speech dynamics; this typically leads to improved performance. The speech frames are used for acoustic analysis.

**Acoustic models.** In order to analyze the speech frames for their acoustic content, we need a set of *acoustic models*. There are many kinds of acoustic models, varying in their representation, granularity, context dependence, and other properties.

**Acoustic analysis and frame scores.** *Acoustic analysis* is performed by applying each acoustic model over each frame of speech, yielding a matrix of *frame scores*. Scores are computed according to the type of acoustic model that is being used. For template-based acoustic models, a score is typically the Euclidean distance between a template's frame and an unknown frame. For state-based acoustic models, a score represents an *emission probability*, i.e.,

the likelihood of the current state generating the current frame, as determined by the state's parametric or non-parametric function.

**Time alignment.** Frame scores are converted to a word sequence by identifying a sequence of acoustic models, representing a valid word sequence, which gives the best total score along an *alignment path* through the matrix1, as illustrated in the Figure. The process of searching for the best alignment path is called *time alignment*. An alignment path must obey certain *sequential constraints* which reflect the fact that speech always goes forward, never backwards. These constraints are manifested both within and between words. Within a word, sequential constraints are implied by the sequence of frames (for template-based models), or by the sequence of states (for state-based models) that comprise the word, as dictated by the phonetic pronunciations in a dictionary, for example. Between words, sequential constraints are given by a grammar, indicating what words may follow what other words.

Time alignment can be performed efficiently by *dynamic programming*, a general algorithm which uses only local path constraints, and which has linear time and space requirements. (This general algorithm has two main variants, known as *Dynamic Time Warping* (DTW) and *Viterbi search*, which differ slightly in their local computations and in their optimality criteria.)

In a state-based system, the optimal alignment path induces segmentation on the word sequence, as it indicates which frames are associated with each state. This segmentation can be used to generate labels for recursively training the acoustic models on corresponding frames.

- **Word sequence.** The end result of time alignment is a *word sequence* the sentence hypothesis for the utterance. Actually it is common to return several such sequences, namely the ones with the highest scores, using a variation of time alignment called *N-best search* (Schwartz and Chow 1990). This allows a recognition system to make two passes through the unknown utterance: the first pass can use simplified models in order to quickly generate an N-best list, and the second pass can use more complex models in order to carefully rescore each of the N hypotheses, and return the single best hypothesis.

### 2.3.3. Speech Signal Representation

Speech is a sound wave of continuously varying air pressure. Representation of this air pressure variation in a form accessible to computers is fundamental in the analysis of speech. For that purpose the speech signal is first converted into an electrical signal using a microphone and then transformed into a discrete signal-using analog to digital converters (Yaung, Gunnar, Mark, Thomas, Dan K., Xunying, Gareth, Julian, Dave, Dan P., Valtcho, and Phil 2006).

For speech recognition applications the signal undergoes further processing stages. Most processing algorithms perform spectral analysis over a window

containing a number of speech samples. The window is then moved in increments of a few milliseconds. These windows are called *analysis* frames (Solomon 2005). Short time Fourier transforms are applied on analysis frames to obtain what are known as the *spectral coefficients* of the speech signal for the frame. The spectral coefficients are an alternative form of representing the speech signal in the analysis frame. The resulting coefficients form a *feature vector* for the frame. The feature vectors are used as representatives of the speech signal within that frame and are used in the computation of the acoustic probability mentioned earlier. To capture changes in the signal, a feature vector can additionally contain features of neighboring frame.

Several different ways of extracting the features of speech can be used. Almost all of them are based on the source-filter model of speech production. This speech production model assumes that speech is generated through a process of filtering a source signal. Figure 2.2 shows the basic source filter-model for speech signal.

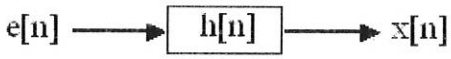


Figure 2.2. Basic source- filter model of speech production (Adapted from (Young, Gunnar, Mark, Thomas, Dan K., Xunying, Gareth, Julian, Dave, Dan P., Valtchev, Phil 2006))

The source signal is provided by a stream of air pushed out of the lungs through the vocal tract. This stream of air vibrates the vocal cords for voiced phonemes or gets to the oral or nasal tract unhindered for unvoiced phonemes. This is modeled by choosing the source signal **e[n]** to be either random noise or a periodic pulse to mimic unvoiced and voiced phonemes, respectively.

The stream of air that comes out of the lungs through the vocal tract is then modified by the oral and/or nasal tract. To articulate a particular phoneme, these tracts change their shape of geometry. This is achieved by changing the relative position of the various articulatory organs in the vocal and nasal tract, which include the tongue, the lips, the velum and the oral cavity. The effect of this modification is to change the resonance frequencies, which are referred to as *formant frequencies*, of the nasal and oral tracts. The formant frequencies of each phoneme are different and hence the final articulation of each phoneme will have its own formant frequency profile. In the source filter model of speech production, the last process is modeled through a filter. The source signal is fed to a filter and the output of this filter **x[n]** will be a realization of the desired phoneme (or longer speech segments, if longer durations are taken and the filter parameters are time dependent). Obviously, each phoneme will have its own filter parameters **h[n]** to reflect the particular formant frequencies corresponding to the shape of the vocal and nasal tracts.

The most obvious way of modeling each phoneme is, therefore, to choose the parameters of the filter **h[n]** as the feature vectors. There are several candidate filter models, the most common ones of which include, linear prediction coefficients, line spectral frequencies, cepstral coefficients and mel frequency cepstral coefficients.

### 2.3.4. Approaches to the Development of Speech Recognition

Broadly speaking, there are three approaches to speech recognition namely acoustic-phonetic approach, pattern recognition approach and artificial intelligence approach (Juang and Rabiner 1993).

#### 2.3.4.1. Acoustic-Phonetic Approach

It is based on the theory of acoustic phonetics that postulates that there exists finite, distinctive phonetic units in spoken language and that the phonetic units are broadly categorized by a set of properties that are manifested in the speech signal, or its spectrum, over time. Even though the acoustic properties of phonetics units are highly variable, both with speakers and neighboring phonetic units (the so-called articulation of sounds), it is assumed that the rules governing the variability are straightforward and can readily be learned and applied in practical situations.

The Acoustic-Phonetic approach assumes that the phonetic units are broadly characterized by as set of features such as voiced/unvoiced and pitch. These

features are extracted from speech signal and are used to segment and label the speech. The process of recognition in this approach involves three steps (Rabiner and Juang 1993): The first step 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 phase is 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 phone lattice characterization of the speech.

The last step attempts to identify a valid word (or string of words) from the sequence of phonetic labels produced by segmentation and labeling.

The Acoustic-Phonetic approach has not been widely used in most of the ASR systems. Rabiner and Juang (1993) mentioned four major problems that account for the lack of success of this approach in speech recognition systems. These are:

1. It requires extensive knowledge of acoustic properties of phonetic units.

2. The choice of features is mostly based on adhoc considerations.

3. The design of sound classifiers is also not optimal.

4. No well defined, automatic procedure exists for tuning the method on real, labeled speech.

### 2.3.4.2. Artificial Intelligence Approach

It is a hybrid of acoustic phonetic approach and pattern recognition approach in that it exploits ideas and concepts from both methods. The AI approaches tries to mechanize the recognition procedure according to the way a person applies his/her intelligence in visualizing, analyzing, and finally making a decision on the measured acoustic features. The basic idea of AI approach is to compile and incorporate information drawn from a variety of knowledge sources into the system. Thus, for example the AI approach to segmentation and labeling would be to augment the generally used acoustic knowledge with the other high level information sources, like phonemic, lexical, syntactic, semantic and even pragmatic knowledge. That is, the Artificial Intelligence approach incorporates knowledge about the world and the background of the speech into the ASR system. According to Rabiner and Juang (1993:13), among the techniques used within this class of methods are:

1. The use of an expert system for segmentation and labeling such that this crucial and most difficult step can be performed taking more and other knowledge into account rather than just the acoustic information used by pure Acoustic-Phonetic methods;

2. Learning and adapting over a period of time;

3. The use of neural networks for learning the relationship between phonetic events and all known input, as well as to discriminate between similar sound classes.

### 2.3.4.3. Pattern-Recognition Approach

The most known and well performing method for speech recognition is the pattern recognition approach. In the pattern recognition approach, the speech patterns are used directly without explicit determination of phonetic feature and segmentation. It requires no explicit knowledge of speech. This approach has two steps, namely, training of speech patterns based on some generic spectral parameters and recognition of patterns via pattern comparison. Speech knowledge is brought into the system via the training procedure. In the pattern-recognition approach, all acoustic realizations of units, words and sentences are considered as patterns consisting of sequences of feature vectors. Sentence recognition is, therefore, accomplished by performing pattern matching at unit, word and sentence levels in an integrated manner. This approach is the most common one for three basic reasons (|Rabiner and Juang, 1993:15):

1. ***Simplicity of use***: The method is easy to understand, and it is rich in mathematical and communication theory justification for individual procedures used in training and decoding, and it is widely used and understood.

2. ***Robustness and invariance to different speech vocabularies, users, feature sets, pattern comparison algorithms and decision rules***: This property makes the algorithm appropriate for a wide range of speech units (ranging from phoneme like units all the way through words, phrases and sentences), word vocabularies, background environments, transmission conditions, etc.

3. ***Proven high performance***: It has already been shown that the pattern recognition approach provides high performance than the other approaches.

The most successful and popular method of pattern recognition approach in the area of speech recognition is the Hidden Markov Model (HMM) (Young, Gunnar, Mark, Thomas, Dan K., Xunying, Gareth, Julian, Dave, Dan P., Valtchev, Phil 2006). An HMM is a collection of states connected by transitions. An N-state Markov Model is completely defined by a set of N states forming a finite state machine and an N x N stochastic matrix defining transitions between states whose elements aij = P(state j at time t | state i at time t-1) are the transition probabilities. Its output symbols are probabilistic, and all symbols are possible at each state. Therefore, each state or transition is associated with a probability distribution of all possible symbols. An HMM is composed of a non-observable "hidden" process (a Markov chain) and an observation process which links the acoustic vectors extracted from the speech signal to the states or transitions of the "hidden" process. In that sense, an HMM is a doubly stochastic process. The mathematics framework of the HMM method enables us to combine modeling

of stationary stochastic process (for the short time spectra) and the temporal relationship among the processes, (via a Markov chain) together in a well defined probability space. Another advantage of HMM comes from the fact that it is relatively easy and straight forward to train a model from a given set of labeled training data.

Flexibility is also an attractive feature of the basic HMMs (Young, Gunnar, Mark, Thomas, Dan K., Xunying, Gareth, Julian, Dave, Dan P., Valtchev, Phil 2006). It is manifested in three aspects of the model, namely: observation distributions, model topology and decoding hierarchy. We can develop either discrete HMMs or continuous HMMs. In discrete HMMs, distributions are defined on finite spaces while in continuous HMMs; distributions are defined as probability densities on continuous observation spaces. We do also have different alternatives of HMM topologies with different number of states. It is also possible to build HMMs that can decode speech in various hierarchies that range from phones to sentences.

These strengths have made HMMs the predominant method in current automatic speech recognition technology and research. An HMM has the following five basic parameters (Rabiner and Juang 1993):

1. N, the number of states in the model. We denote the set of all possible states as

   S = {S1, S2, . . . . , SN}, the state at time t as qt.

2. M, the number of distinct observation symbols per state, i.e., the discrete alphabet size of the output set. We denote the set of all possible output symbols as V = {v1, v2, ....,vM}, the output symbol at time t as $O_t$. The sequence of observed symbols is denoted as

$$O = O_1O_2....O_T.$$

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

$$a_{ij} = P[q_{t+1} = j \mid q_t = i],\ 1 \leq i, j \leq N$$

4. The observation symbol probability distribution, B = $\{ b_j(k)\}$, in which

$$b_j(k)=P[o_t = v_k \mid q_t =j],\ 1 \leq k \leq M,$$

defines the symbol distribution in state j, j=1,2, . . . . , N.

5. The initial state distribution $\pi = P[q_i = i],\ 1 \leq i \leq N$

The HMM model in the figure below is a model of five emitting states that are numbered from 2 to 6 and output probability distributions associated with them. States number 1 and 7 are non-emitting states and serve only to join models together.

The arrows from one state to the other and the indexed letter 'a' indicate the transition lines and their probabilities respectively. For example $a_{12}$ means the

probability of transition from state 1 to state 2 and $a_{22}$ means the probability of looping in state 2.

In a first order hidden Markov model, there are two assumptions. The first is the Markov assumption. It states "*the probability that the Markov chain is in a particular state at time t + 1 depends only on the state of the Markov chain at time t and is conditionally independent of the past*". The second is "*the output-independence assumption according to which the probability that a particular symbol will be emitted at time t depends only on the state at the time and is conditionally independent of the past*". Although these assumptions severely limit the memory of the first-order hidden Markov models, they reduce the number of parameters and also make learning and decoding algorithms extremely efficient (Young, Gunnar, Mark, Thomas, Dan K., Xunying, Gareth, Julian, Dave, Dan P., Valtchev, Phil 2006).

There exist three problems when constructing HMM. For an HMM model to be useful in building speech recognizers, three fundamental problems must be solved (Rabiner and Juang, 1993). These problems are

Problem 1: Given the model parameters, how do we compute the probability of a particular output sequence?

Problem 2: Given the model parameters, how do we find the most likely sequence of hidden states which could have generated a given output sequence?

Problem 3: Given an output sequence, how do we find the most likely set of state transition and output probabilities? In order to deal with these problems, HMM possesses elegant and efficient algorithms (Young, Gunnar, Mark, Thomas, Dan K., Xunying, Gareth, Julian, Dave, Dan P., Valtchev, Phil 2006); namely,

1. The forward algorithm
2. The Viterbi algorithm
3. The Baum-Welch algorithm

Problem1 is a problem of evaluating how well a given model matches a given observation sequence. To solve this problem, the forward algorithm is used. The forward algorithm is an inference algorithm for HMMs which computes the posterior marginal of all hidden state variables given a sequence of observations/emissions i.e. it computes, for all hidden state variables, the distribution. To solve problem 2; that is, to find the single best state sequence, Q = $\{q_1, q_2, \ldots, q_T\}$ for the given observation sequence O=$\{O_1, O_2, \ldots, O_T\}$, the Viterbi algorithm is used. The Viterbi algorithm chooses the best state sequence that maximizes the likelihood of the state sequence for the given observation sequence.

Problem 3 is solved by using the Baum-Welch re-estimation algorithm. Baum-Welch re-estimation algorithm computes maximum likelihood estimates and posterior mode estimates for the parameters (transition and emission probabilities) of an HMM, when given only emissions as training data.

## 2.4. The HMM and HTK

The Hidden Markov Model (HMM) is a well known stochastic model most widely used for estimating the acoustic probability P(S/W). HMM uses states and transitions to describe a class of random variables. The output sequence of the process is governed by the probability of each state to produce an output symbol and by the probability of making a transition from one state to another. Each state, therefore, has a probability distribution to produce the possible output symbols and another probability distribution to describe the likelihood of taking a transition to another state.

### 2.4.1. The Hidden Markov Model Toolkit

HTK is a toolkit for building Hidden Markov Models (HMMs). HMMs can be used to model anytime series and the core of HTK is similarly general-purpose (Young et.al. 2002). However, HTK is primarily designed for building HMM-based speech processing tools, in particular recognizers. HTK consists of a set of library modules and tools available in (C) source code. The tools provide sophisticated facilities for speech analysis, HMM training, testing and results analysis. The HTK

tools are prepared for all the four processing steps involved in building a sub-word based continuous speech recognizer. These four main phases or steps are: data preparation, training, testing and result analysis.

Data preparation includes speech recording, preparation and formatting of the associated transcriptions that use sub-word or word labels and parameterization of the training and test speech data. HTK provides the required modules for this data preparation. It also provides tools to assist in constructing the pronunciation dictionary, which may be considered to belong to data preparation (Young, Gunnar, Mark, Thomas, Dan K., Xunying, Gareth, Julian, Dave, Dan P., Valtchev, Phil 2006).

HTK enables both methods of model initialization that are mentioned above. The HTK training tools are used to estimate the parameters of a set of HMMs using training utterances and their associated transcriptions. It uses the Baum-Welch re-estimation procedure for this purpose. Although HTK gives full support for building whole word HMM systems, the bulk of its facilities are focused on building sub-word systems.

Speech utterance can be transcribed using the HTK recognition tool that performs Viterbi-based speech recognition. It takes as input, a network (language model) describing the acceptable word sequences, a pronunciation dictionary that defines 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 definition to each sub-word instance. Recognition can then be performed on either a list of stored speech files, or on direct audio input. The discussion of the toolkit is mainly based on the handbook prepared by (Young et.al. 2002).

### 2.4.2. Basics of HMMs

It has been shown that the a priori probability of a sequence of words P(W) and the a posteriori probability of the sequence given an observation P(W/O) - the two models in equation shown above - are important and are modeled separately. The former is determined by a language model where as the later is given by an acoustic model.

In most state-of-the-art recognition systems, the hidden Markov model (HMM) is used in the acoustic modeling. HMM provides a simple means to estimate the above conditional probabilities on the basis of finite-state Markov chains (Mesfin 2008). It is assumed that a Markov model generates the sequence of observed speech parameter vectors corresponding to each word.

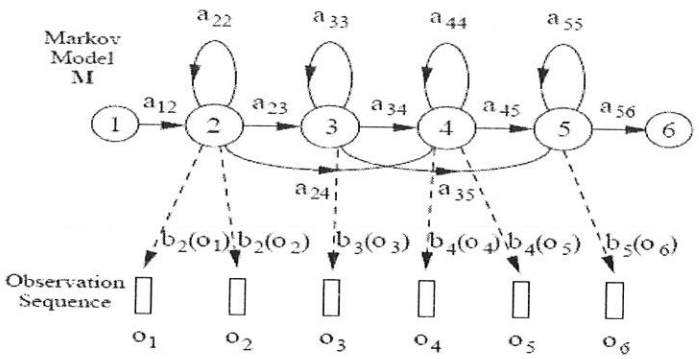


Figure 2.3. A Markov Generation Model (Adapted from (Young, Gunnar, Mark, Thomas, Dan K., Xunying, Gareth, Julian, Dave, Dan P., Valtchev, Phil 2006))

An HMM contains a Markov distribution for transitions across different states, and includes a probability density function at each state that models the probability of the output symbols possible at that state. It is a doubly stochastic state machine that can be fully described by the triple {S, A, B}. Here, S is the initial state probability, A = {aij} is the state-transition probability set, and B = bj(Ot) is the emission probability distribution (Hamaker 2002) cited by (Zegeye 2003).

## 2.5. Types of ASR

ASR products have existed in the marketplace since the 1970s (Muhirwe 2005). However, early systems were expensive hardware devices that could only recognize a few isolated words (i.e. words with pauses between them), and needed to be trained by users repeating each of the vocabulary words several

times. The 1980s and 90s witnessed a substantial improvement in ASR algorithms and products, and the technology developed to the point where, in the late 1990s, software for desktop dictation became available 'off-the-shelf' for only a few tens of dollars (Muhirwe 2005). From a technological perspective it is possible to distinguish between two broad types of ASR:

- *Direct Voice Input (DVI) and*
- *Large Vocabulary Continuous Speech Recognition (LVCSR)*

DVI devices are primarily aimed at voice command-and-control, whereas LVCSR systems are used for form filling or voice-based document creation. In both cases, the underlying technology is more or less the same. DVI systems are typically configured for small to medium sized vocabularies (up to several thousand words) and might employ word or phrase spotting techniques. Also, DVI systems are usually required to respond immediately to a voice command. LVCSR systems involve vocabularies of perhaps hundreds of thousands of words, and are typically configured to transcribe continuous speech. Also, LVCSR need not be performed in real-time - for example, at least one vendor has offered a telephone-based dictation service in which the transcribed document is e-mailed back to the user. Speech recognition systems can be categorized based on different parameters (Mesfin 2008), some of the more important of which are shown in Table below.

| Parameters | Range |
|---|---|
| Speaking Mode | Isolated Words vs. Continuous speech |
| Speaking Style | Read Speech vs. Spontaneous Speech |
| Enrollment | Speaker-dependent vs. Speaker-independent |
| Vocabulary Size | Small (less than 20 words identified) vs. Large (greater than 20,000 words identified) |
| Signal to Noise Ratio (SNR) | High (>30dB) vs. Low (<10dB) |

Table 2.1. Typical parameters used to characterize the capability of speech recognition systems

## 2.6. Application of Automatic Speech Recognition

Speech recognition systems are applied in different application domains. Some of the most common application areas of speech recognition systems include dictation systems, command and control systems, telephony systems, as an assistive technology for disabled people like the handicapped that cannot use keyboard or any pointing devices, audio based information retrieval systems etc. Dictation system includes medical transcriptions, legal and business dictation, and general word processing. Command and control systems use speech input to perform functions and actions. Telephony systems allow callers to speak

commands instead of pressing buttons to dial a number. In general, apart from being a natural way of interfacing with machines (like PCs, Telephones, Television etc), ASR renders the following advantages:

- it helps to speed up inputting information (much faster than using the ubiquitous, keyboard even for the fastest possible typists);
- it avoids pains related to typing (like repetitive stress injury);
- using ones voice, the hands and the eyes are free and movement is unconstrained.

Speech recognition applications vary depending on the vocabulary size, the quality of the microphone, the number of users, and the tolerance for error of the users. A few examples include (Mesfin 2008:21):

- ***Hands-free control of machinery:***
  - ✓ Small vocabulary,
  - ✓ Speaker-independent,
  - ✓ High-quality microphone (often a headset microphone),
  - ✓ Very low error tolerance (error tolerance can be increased with verbal feedback).

- ***Automatic telephone dialing***

  - ✓ Small vocabulary,
  - ✓ Mixture of speaker-independent (digits) and speaker-dependent (names),
  - ✓ Low-quality microphone (telephone handset),
  - ✓ Moderate error tolerance (if the system asks for confirmation before dialing).

- ***Telephone access to databases***

  - ✓ Moderate vocabulary,
  - ✓ Speaker-independent,
  - ✓ Low-quality microphone,
  - ✓ High error tolerance.

- ***Word processing***

  - ✓ Large vocabulary,
  - ✓ Speaker-dependent,
  - ✓ High-quality microphone,
  - ✓ High error tolerance.

## 2.7. Challenges in Speech Recognition

The problem of automatically recognizing speech with the help of a computer is a difficult problem, and the reason for this is the complexity of the human language. Humans use more than their ears when listening; they use the knowledge they have about the speaker and the subject. Words are not arbitrarily sequenced together, there is a grammatical structure and redundancy that humans use to predict words not yet spoken. Furthermore, idioms and how we 'usually' say things makes prediction even easier (Jurafsky 2009).

In ASR we only have the speech signal. We can of course construct a model for the grammatical structure and use some kind of statistical model to improve prediction, but there are still the problem of how to model world knowledge, the knowledge of the speaker and encyclopedic knowledge. We can, of course, not model world knowledge exhaustively, but an interesting question is how much we actually need in the ASR to measure up to human comprehension (Markus 2003). Some of the factors affecting ASR are:

- ✓ Body language:
- ✓ Noise
- ✓ Difference between spoken language and written language
- ✓ Continuous speech
- ✓ Channel variability
- ✓ Speaker variability

- Realization
- Speaking style
- The sex of the speaker
- Anatomy of the vocal tract
- Speed of speech
- Dialect
  - Regional dialect
  - Social dialect

✓ Amount of data and search space

✓ Ambiguity

- Homophones ambiguity
- Word boundary ambiguity

The focus of this research is on pronunciation variability, which is mentioned above as speaker variability, that is a factor for affecting the Recognition System and is one of the challenges in ASRS. However, these should be taken as a serious point of the Recognizer efficiency improvement. Also Pronunciation variation is very wide concepts that need to be considered while developing ASR for Amharic language too.

# CHAPTER THREE

# THE AMHARIC LANGUAGE

## 3.1. Language and Linguistic Fundamentals

"Whatever else people do when they come together – whether they play, fight, do tasks, or make automobiles-they talk", described by Fromkin (2003:3). We live in the world of language. We talk to our friends, our associates, our wives and our husbands, our lovers, our teachers, our parents, our rivals, and even our telephone, and everyone responds with more talk. Accordingly, language can be taken as an instrument/tool for communication. Formally we can define language as set of variables, attributes, expressions and all actions having given names (Solomon 2001). All languages have their own inventory of speech sounds that are combined to form syllables and words. The words themselves are combined to form phrases and sentences, leveraging, in effect, the expressive capabilities of relatively small number of speech sounds. This makes speech sounds the very essence in any language related studies.

The possession of language, perhaps more than any other attribute, distinguishes humans from other animals. To understand our humanity, one must understand the nature of language that makes us human. According to the philosophy expressed in myths and religions of many people, language is the source of human life and power. When we know a language, we can speak

and be understood by others who know that language. This means we have the capacity to produce sounds that signify certain meanings and to understand or interpret the sounds produced by others (Referring only to normal-hearing individuals). Knowledge of a language enables us to combine words to form phrases, and phrases to form sentences. We cannot buy a dictionary of any language with all its sentences, because no dictionary can list all the possible sentences. Knowing a language means being able to produce new sentences never spoken before and to understand sentences never heard before. Understanding a sentence involves analysis at many levels. To begin with, we must comprehend the individual speech sounds we hear.

In linguistic theory, the analysis and description of linguistic phenomena are usually organized into several distinct levels. The different sounds used by a language are described at the level of *phonology*. The writing system is described at the level of *orthography*. *Morphology* describes the formation and inflection of individual words. *Syntax* describes the ordering of words and their combination into phrases and sentences. *Semantics* analyzes the meaning of individual words (*lexical semantics*) and the meaning of phrases and sentences (*compositional semantics*). How words and phrases are actually used to make things happen is the level of *pragmatics*. How people and things are introduced as topics and subsequently referred to in later utterances is the level of *discourse*.

When dialogues between humans and computers are more natural, the ASR must handle more conversational speech. Conversational speech is harder for ASR systems to recognize correctly, because of increased co-articulation and pronunciation variability, as well as less predictable language usage. Weintraub et al.(1996) showed that a spontaneous speaking style is harder to recognize; when the same exact word sequences were recorded in a truly spontaneous, acted spontaneous, and read style, the ASR system performed much worse on spontaneous speech compared with the other two styles. Ideally, speech recognizers should handle these diverse speaking styles, (e.g. spontaneous speech, hyper-articulated speech, accents, dialects, and speech from users with different mother tongues). This kind of variation in user input is difficult to model and this is not solved for in current state-of-the-art recognizers (Ingunn and Eric 2003).

Therefore, to enable computers to *process language as skillfully as human do* will signal the arrival of truly intelligent machines which understood the natural language (). This will lead us to review the fundamental knowledge to the areas of linguistic that are related to the development of ASR systems.

***Phonetics*** is the study of speech sounds used in language of the world. It is concerned with the sounds of languages, how these sounds are articulated and how the hearer perceives them. Phonetics is related to the science of acoustics

in that it uses much of the techniques used by acoustics in the analysis of sound (Abraham 2011).

The anatomy of the human speech production system is shown in Figure 3.1. The vocal apparatus comprises three cavities (Abraham 2011): nasal, oral, and pharyngeal. The pharyngeal and oral cavities are usually grouped into one unit referred to as the vocal tract, and the nasal cavity is often called the nasal tract. The vocal tract extends from the opening of the vocal folds, or glottis, through the pharynx and mouth to the lips (shaded area in Figure). The nasal tract extends from the velum (a trapdoor-like mechanism at the back of the oral cavity) to the nostrils.

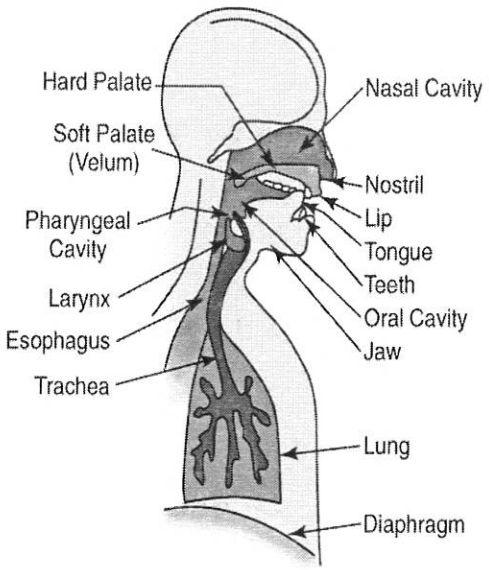


Figure 3.1. The human speech production system (Adapted from (Adami, 2005)

From their acoustic features sounds or phones can be categorized into consonants and vowels. Phones are physical sounds and one of the most common sub-word recognition units used in the development of automatic speech recognizers.

As pointed out by (Peter 2001), there are three sub-disciplines of phonetics that study the difference in features of speech sounds. These are:

1. ***Articulatory Phonetics:*** The study of the production of speech sounds. It is the study of articulators in the process of the production of speech sounds. Articulators are parts of the local tract where we have a large and complex set of muscles. The muscles change the shape of the articulators enabling them to modify the flow of air that passes from the chest through the mouth and nostrils into the atmosphere. It is this modification that makes speech sounds different from each other in their acoustic features.

2. ***Acoustic Phonetics***: The study and analysis of the physical production and transmission of speech sounds. Speech sounds, like sounds in general, are transmitted through the air as small, rapid variations in air pressure that spread in longitudinal waves from the speaker's mouth and can be heard, recorded, visualized and measured. Differences between individual speech sounds are directly reflected as differences in either one or several or all of the sound parameters, like tone, stress, duration, pitch, loudness and quality of the speech waves. By dealing with the study and description of the acoustic

properties of individual speech sounds, acoustic phonetics is the immediate link between articulatory phonetics and speech perception. It is also important for applications in the fields of signal processing and speech technology, like ASRS.

3. ***Auditory Phonetics***: The study of the perception of speech sounds. Just as articulatory phonetics involves the understanding of the anatomy of the human speaking system, auditory phonetics involves the understanding of the human hearing system. This means, auditory phonetics deals with the understanding of the anatomy and physiology of the human ear and brain. An ideal automatic speech recognizer is on that recognizes all the speech that the human auditory system recognizes. That is why some knowledge generated by auditory phonetics is applied in the development of ASR.

***Phonology*** is the study of the sound patterns of a language. It describes the systematic way in which sounds are differently realized in different contexts, and how this system of sounds is related to the rest of the grammar. Phonology is concerned with how sounds are organized in a language. It endeavors to explain what these phonological processes are in terms of formal rules (Jurafsky 2009).

A speech recognition system needs to have a pronunciation for every word it can recognize, and a text-to-speech system needs to have a pronunciation for every word it can say. Modeling pronunciation would be much simpler if a

given phone was always pronounced the same in every context. In reality this is not the case, for example the phone[t] is pronounced very differently in different phonetic environments. In general definition phonology is the area of linguistics that describes the systematic way that sounds are differently realized in different environments, and how this system of sounds is related to the rest of the grammar (Abraham 2011).

Part of the phonological study of a language involves analyzing phonetic transcriptions of speech made by native speakers and trying to deduce what that underlying phonemes are and what the sound inventory of the language is. Looking for minimal pairs forms part of the research in studying the phoneme inventory of a language. HMM-based large vocabulary ASRSs model an HMM for every phoneme or group of phonemes that are in the sound inventory of the language. The required pronunciation dictionary of ASRSs is also prepared in terms of phones or groups of them like syllables. Phones can be described by how they are produced articulatorily by the vocal organs; are consonants are defined in terms of their *place* and *manner* of articulation and *voicing,* vowels by their *height* and *back-ness.* A phoneme is the smallest meaning distinguishing unit or can be defined as a generalization or abstraction over different phonetic realizations.

***Morphology*** is the study of words and word structure of a given language. Meaningful components of a word, word formation, shape of a word and

structure are studies in morphology. It deals with how words and subparts of words are put together from their smaller parts and the rules governing this process. The elements that are combined to form words are called morphemes. A morpheme is often defined as the minimal meaning-bearing unit in a language. So for example the word *fox* consists of a single morpheme (the morpheme fox) while the cats consists of two: the morpheme cat and the morpheme –s. As the example shows, it is often useful to distinguish two broad classes of morphemes: *Stems* and *affixes*. The exact detail of the distinction vary from language to language, but intuitively, the stem is the "main" morpheme of the word, supplying the main meaning, while the affixes add "additional" meanings of various kinds (Jurafsky 2009).

***Syntax*** is the study of the structural relationship between words. If words are the foundation of speech and language processing, syntax is the skeleton. It deals with formal relationships between words, attempts to describe what is grammatical in a particular language in terms of rules. Syntactic knowledge of a language is given to an ASRS as its language model. The main purpose of the syntax component of an ASRS is to constrain the number of word sequences to be dealt with in the recognition process and to predict or insert poorly recognized words. For example, statistical language models estimate the probability of word sequences which are possible sentences in the language. Statistical language models give no clear-cut dichotomy between grammatical and ungrammatical sentences of the language. They rather give higher

probability for more frequent phrases and lower probability for less frequent ones. A language model may also provide aspects of the semantics and pragmatics of a language for the ASRS (Jurafsky 2009).

## 3.2. Overview of the Amharic Language

Amharic (also known as Abyssinian, Amarinya, Amarigna, and Ethiopian) is the working language of Federal Republic of Ethiopia. Since the 13$^{th}$ century has been the language of the court and dominant population in Highland Ethiopia (Amanda and Rachel 2011). The language of Amharic is spoken in the Ethiopian government, court system, and on all official documents. Amharic is predominately spoken by peoples found in all regional states of Ethiopia. According to the 2007 census Amharic is the most commonly spoken language in Ethiopia. Amharic speakers encompass 32.7% of Ethiopia's population. Amharic is also spoken by 40,000 people in Israel as well as people in Egypt and Sweden (Amanda and Rachel 2011).

The Amharic alphabet consists of thirty-three basic characters, each of which has six additional modified characters. The modified characters represent the basic sound of the symbol augmented with a vowel. Thus, the main table of the traditional Amharic syllabary, appears as characters set in thirty-four rows and seven columns. Languages using such a scheme have been termed to use an "Abugida" instead of an alphabet. Abugida is a term coined by Peter Daniels for a script whose basic signs denote consonants augmented with a vowel and

where consistent modifications of the basic sign indicate augmentation of other vowels. The term "Abugida" is derived from the first four characters of one type of ordering of the Ge`ez script. The Amharic script is an FidelAbugida, although the vowel modifications in Amharic are not entirely systematic (Mesfin, 2008).

As a result of using Fidel/Abugida, each concatenated consonant-vowel syllable in the Amharic language has its own corresponding orthographic symbol. A few syllables, however, are exceptions to this rule. These syllables are represented by more than one symbol in the character set. These redundant orthographic symbols used to make differences in sound realization as well as meaning in old Amharic. In modern Amharic, these differences do not exist.

Another implication of the use of Fidel/Abugida is that there are a large number of characters in the language. This makes encoding Amharic using a computer keyboard difficult. Using current Amharic text encoding programs and typewriters, it routinely takes 2 or 3 keystrokes to type in a single character. It is not unheard of to find characters that take 4 keystrokes on some of the Amharic text encoding programs. It is instructive to compare this with English that requires only a single keystroke for all its characters except for certain special characters and capital letters in which case only two key strokes are necessary.

## 3.3. Linguistic Features of Amharic Language

### 3.3.1. Phonology

In the International Phonetic Alphabet translation of Amharic sounds Symbols in parentheses represent deviations from standard IPA symbols. Amharic includes glottalized series of consonant phonemes, which is characteristic of the sound system. Syllable structure is represented as CVCC (Solomon 2001). Consonant clusters will not appear in initial position. Stress may occur on each syllable, but the last syllable tends to be unstressed (Solomon 2001).

| Manner of Articulation | Airstream and modifications | Place of Articulation | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Bilabial | Labio-dental | (Denti-)[1] Alveolar | Post-alveolar | Palatal | Velar | Glottal |
| | | vl vd | vl vd | vl vd | vl vd | vl vd | vl vd | vl |
| Stop | Pul Simp<br>Pul Lab<br>Glot Simp<br>Glot Lab | p b<br>p' | | t d<br>t' | | | k g<br>kʷ gʷ<br>k'<br>k'ʷ | ʔ |
| Fricative | Pul Simp<br>Glot Simp | | f (v[2]) | s z<br>s' | ʃ ʒ | | | h |
| Affricate | Pul Simp<br>Glot Simp | | | | tʃ dʒ<br>tʃ' | | | |
| Nasal | Pul Simp | m | n | | | ɲ | | |
| Central Approximant | Pul Simp | w | | | | j | | |
| Lateral Approximant | Pul Simp | | l | | | | | |
| Trill | Pul Simp | | r | | | | | |

Table 3.1. Consonants of the Amharic Language (adapted from (Derib 2011)).

| | Front | Central | Back |
|---|---|---|---|
| High | i | ɨ | u |
| Mid | e | ə | o |
| Low | | a | |

Table 3.2. Vowels of Amharic Language (Adapted from (Derib 2011)).

#### 3.3.2.1.1. Writing System

Amharic has its own writing system, a semi-syllabic system. There is no agreed translation of Amharic symbols to Roman characters (used in English). There are 33 consonant symbols that have seven variations. Variations are according to the vowel that is coupled with the consonant (Baye 2010).

In Amharic, the consonant sounds /p/ፕ/, /t/ት/, /k/ክ/, and /s/ስ/ can be produced as ejectives. Additionally some predictable patterns of Amharic speakers are that final consonants are often devoiced or deleted, fricatives may become stops, stops may become fricatives, and vowels are often shortened, lowered, or raised which can also be taken as a reason for the variation of pronunciation.

### 3.3.3. Morphology

The typical clause order in Amharic is ***noun + object + verb*** (Amanda and Rachel 2011).

- ***Nouns:*** may denote gender, number, definiteness, case, and direct object status by affixes prefixes and suffixes, predominately suffixes. Amharic nouns may have a masculine or feminine gender. Suffixes are added to denote a masculine or feminine noun gender. Some nouns may have both masculine and feminine gender, while other nouns may only have one gender. The feminine gender is used to indicate female as well as smallness. Plurals are indicated by the suffix *–occ/sh/*. Affixes are added in the following order: *gender, number, definiteness, case, and direct object status*.
- ***Pronouns:*** Amharic is a pro-drop language. Sentences with no emphasized element do not require independent pronouns. The verb denotes the person, number, and gender. Object pronoun suffixes are affixed to verbs and indicate person, number, and gender of the object of a verb. Possessive suffixes are affixed to nouns to indicate possession.
- ***Verbs:*** are derived from roots and affixes to inflect person, number, gender, aspect, mood, voice, and polarity are added. Verbs agree with their subjects. Verb agreement with objects is optional. Verbs are placed at the end of the sentence.
- ***Adjectives:*** are predominately derived from nouns, verbs, and other parts of speech.

The first problem of different symbols arising from the similar speech waveforms is not as pronounced in Amharic as in other languages such as English. This is a direct result of the systematic representation of speech sounds by a consistent set of symbols in the Amharic Fidel/Abugida. As a result mapping speech waveforms to orthographic symbols is not as difficult as (at least in theory) other languages. The Amharic orthography consists of 276 distinct symbols excluding the twenty numeral symbols and the eight punctuation marks. The main task of speech recognition is concerned with identifying distinct speech units, thus as a decent approximation the writing of Amharic the redundant symbols may be taken out of the set of symbols in the speech recognizer without losing the essential understanding in the text. This results in a total of only 233 graphemes (syllabic characters) in the language (Mesfin 2008).

Making speech technology based applications more widespread has several consequences for the demands on ASR systems. To increase modeling capacities of ASR systems for Amharic language also need to manage the looser articulation of continuous speech. All the above-mentioned linguistic features also bring variability in pronunciation of the words in the language, which should be considered, while the development of Automatic speech Recognition for Amharic language.

# CHAPTER FOUR

# PRONUNCIATION VARIATION IN AMHARIC SPEECH RECOGNITION

## 4.1. Introduction

In this thesis, the researcher's hypothesis was pronunciation variability has an effect on the development of ASR for Amharic language also the performance can be improved by using the direct data-driven approach. Here in this chapter the discussion is about the effects of pronunciation variation in Automatic Speech Recognition development for Amharic language, which is an issue and has contribution for the improvement in the performance of ASR system for Amharic with limited resources.

## 4.2. Pronunciation Variation

Pronunciation variability is common between natural language speakers. A first major distinction can be drawn between inter-speaker and intra-speaker pronunciation variation. Inter-speaker variation refers to variation in pronunciation of different speakers, whereas intra-speaker variation refers to pronunciation variation of the same speaker. To a large degree inter-speaker variation is caused by anatomical differences between speakers. For example,

male and female speakers and children have different speech characteristics. Inter-speaker variation also exists due to the fact that speakers of the same language may speak different dialects or speak with a different accent (Laver, 1994) cited by (kessen 2002). The accent will depend on factors such as region of origin, socioeconomic background, and level of education, sex and age. Obviously, the speech signal not only conveys the linguistic information (the message) but also a lot of information about the speaker himself: gender, age, social and regional origin, health and emotional state and, with a rather strong reliability, its identity. Apart from the intra- speaker variability (emotion, health, age), it is commonly admitted that the speaker uniqueness results from a complex combination of physiological and cultural aspects (Kessen 2002).

The complex shape of the vocal organs determines the unique "timbre" of every speaker. The larynx which is the location of the source of the speech signal conveys the pitch and important speaker information. The vocal tract can be modeled by a tube resonator. The resonant frequencies (the formants) are structuring the global shape of the instantaneous voice spectrum and are mostly defining the phonetic content and quality of the vowels. "Pronunciation variation" as a term could be used to describe most of the variation present in speech. However, this paper does not deal with the reasons or sources of pronunciation variation, but rather with the implication of pronunciation variations for the development of ASRS for Amharic language which are caused by any reason indicated above. In a language which has much experience in the

development of ASRS, it is widely assumed that pronunciation variation is one of the factors which leads to less than optimal performance in automatic speech recognition (ASR) systems. Pronunciation variation has been identified as a major cause of errors for a variety of automatic speech recognition tasks (McCallester et al. 1998 cited in Kessen 2002). In the baseline system, both the lexicons for training and recognition contain a single phone transcription for each word. This phone transcription is the most likely pronunciation according to the linguistic literature and is called the canonical phone transcription. This degradation in recognition performance is caused by a mismatch between the actual pronunciation of the word and the pronunciation as denoted in the lexicon. This mismatch causes problems both during recognition and training.

## 4.3. Lexical Adaptation Approaches

A lexicon defies the transcription of the words in terms of the acoustic model units of the recognizer. Lexical modification is the most popular way of modeling phonological pronunciation variation. Segmental variation: such as allophonic variation on the target speech. Other types of variation may be better handled at the lexicon level, e.g. insertion deletion, and variation that is present for a group of speakers (e.g. dialects), or is typical for a speaking style. Lexical modeling accommodates longer contexts than acoustic modeling, permitting modeling of syllables and even entire words of phrases (Ingunn 2002).

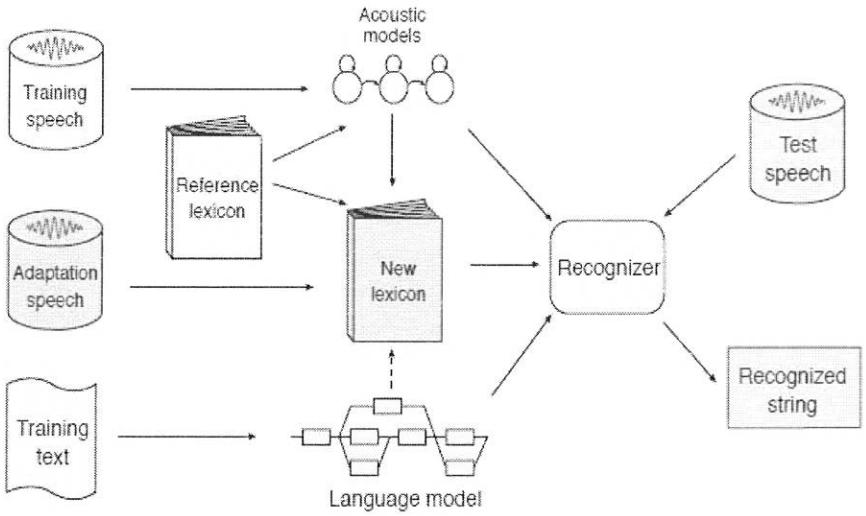


Figure 4.1. Recognition system with adapted lexicon (Adapted from Ingunn 2002)

There are two main directions in finding pronunciation variations, each involving different problems (Ingunn 2002):

1. ***Knowledge based methods***: where we try to find the best pronunciation rules by applying phonetic and linguistic knowledge. The use of linguistically based transformations rules to generate decision trees for alternate pronunciations modeling. Using hand-labeled data is also a kind of knowledge based method, but if the pronunciations found are used in re-transcriptions, the categorizing is less clear cut. The primary advantage of the knowledge-based approach is that it can be applied to all corpora and especially to new words that are not introduced in the ASR system. However, the main problem occurs if the knowledge does not cover the variation we want to model. We may then have too many or too few variations and we may not know how frequent they are.

2. ***Data-driven methods***: where we use databases of real speech to find the variations present. The problem is that the variations based on a given database may give a result too specific for that database. One of the advantages is that we may compute probabilities for the variants, as opposed to the knowledge-based methods. Data-driven methods will model frequently occurring segments better. This might be an advantage, as frequent words will have a larger influence on the WER. Besides, ASR is based on statistics, and the differences and similarities perceived by humans might not be the most useful for ASR. There are also further classifications of data-driven methods:

   - ✓ ***Direct data-driven method:*** derives pronunciation variants depending on pronunciation training databases. No rules or hand-labeled data is used, which will have no prior knowledge of vocabulary other than number of words and boundaries of each word, which is usually present in the orthographic transcription. When an ASR system employs the adapted pronunciation dictionary using a direct data-driven approach, some unseen words might appear during ASR testing. Thus, such a mismatch condition in the pronunciation model between ASR training and testing could degrade the performance of an ASR system (Ingunn 2002).
   - ✓ ***Indirect data-driven method:*** generates pronunciation variants using rules that are automatically derived from data it frequently used. These rules should ideally capture the difference between the reference pronunciation of

a word and the actual pronunciation used by the speakers. An indirect data-driven method investigates pronunciation variability from the speech training data, derives the variant rules, and applies the variant rules in the ASR pronunciation dictionary to compensate for the variability (Ingunn 2002).

Although both knowledge-based and data-derived approaches are used for generating pronunciation variants, they have their drawbacks too. The linguistic literatures, including pronunciation dictionaries are not complete i.e. all pronunciation variations are not described in the linguistic literature or present in pronunciation dictionaries. Furthermore, a knowledge-based approach is subject to discrepancies between theoretical pronunciations and phonetic reality. The major drawback of data-derived approach is it is labor intensive, and therefore expensive. Moreover, manual transcriptions tend to contain an element of subjectivity. Transcriptions made by different transcribers, and even made by the same transcriber, may differ quite considerably (Cucchiarini 1993).

## 4.4. Description of Experiment Setup

The intention of this research is to show the implication and effects of Pronunciation Variation in the development of Automatic Speech Recognition for Amharic. To meet these requirements the recognizer was designed to recognize large vocabulary, continuous speech and was speaker independent. This was implemented using phonemes as base unit.

The development of the experiment and analysis process was performed on Microsoft Windows 7 Ultimate platform using the tools in HTK. We used the Voxforge/Julius package which is Windows plat form for the training of this research experiment. The discussion of the experiment is presented in accordance to the steps that should be followed while building a Large Vocabulary Speaker Independent Speech Recognizer.

### 4.4.1. Major Components

Although there are different kinds of speech recognition systems, most have similar major components. This research mainly focuses on the pronunciation model of ASR. Using data-driven approach to model the pronunciation variation. The major components of an Automatic Speech Recognition system are as follows:

#### 4.4.1. 1. Feature Extraction

The feature extraction component of an ASR system maps the speech waveform into a sequence of feature vectors. This sequence of feature vectors is subsequently used to train acoustic model and decode input speech waveform.

To apply digital signal processing techniques to the speech waveform, the analogue wave form is firstly converted into a digital signal. This is done via sampling and quantization of the waveform. Once the digital signal has been obtained, a variety of techniques can be used to extract features which are useful for the speech classification task. These speech analysis techniques usually assume that the

characteristics of the speech signal are stationary over a short time period, typically of the order of 25 milliseconds. The resulting features are a representation of the speech signal over this short time period. Parameterization is performed not only for size reduction of the original speech signal data but also for pre-processing of that signal that fits into the classification stage. An important property of feature extraction is the suppression of information that is irrelevant for a correct classification such as information about speaker and transmission channel. Currently the most popular features are Mel Frequency Cepstral Co-efficients (MFCC). We have used MFCC for our experiment in developing our research. The speech signal is divided into frames of size 25ms with a frame rate of 10ms.We have used 12 MFCC coefficients and delta.

#### 4.4.1.2. Pronunciation Dictionary

The dictionary provides an association between words used in the task grammar and the acoustic models which may be composed of sub word (phonetic, sysllabic etc,) units. The first step in building a dictionary is to create a sorted list of the required words. The next step is to provide the pronunciation of the words in terms of sub-word units. It specifies the finite set of words that may be output by the speech recognizer and gives, at least, one pronunciation for each. A pronunciation dictionary can be classified as a canonical or alternative dictionary on the basis of the pronunciation it includes (Solomon 2005).

For each word a canonical pronunciation dictionary includes only the standard phone (or other sub-word) sequence assumed to be pronounced in read speech. It does not consider pronunciation variations such as speaker variability, dialect, or Co-articulation in conversational speech. On the other hand, an alternative pronunciation dictionary is a pronunciation dictionary that uses the actual phone (or other sub-word) sequences pronounced in speech. Various pronunciation variations can be included (Fukada et. Al 1999) cited (Solomon 2005). In pronunciation variation research, one is usually confronted with two types of lexicon: as indicated in 5.3., 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 transliteration per word. A multiple pronunciation lexicon contains more than one transliteration variant per word, for some or all of the words in the lexicon. For HTK to compile the acoustic model, we need to make sure that we have at the very least 3 to 5 usage counts for each phone. If there are phones that only have one occurrence, we must add words that use these phones to the prompts file. The phones used were balanced and meet the requirement. Here in both our canonical dictionary and alternative dictionaries, we have checked that we used balanced numbered of phones.

The researcher identified 85 different pronunciation variations for the training corpus and 25 variations for the test corpus. Most of the variations seen were the insertion of the semi-vowels, the insertion of the epenthetic vowel between phones and at the

beginning of a word. There was very observable germination which is common in Amharic language but not considered in this experiment.

There were also words already considered as a different dictionary in the corpus which has implication on the data-driven inputs. For example the word "Si" was read as Si\"ሲ" or Sihe\ "ሲህ" which has an insertion of the 6th order at the end of the phone. But this was transcribed as two phones in the corpus.

#### 4.4.1.3. Language Model

Language model is one of the most important knowledge sources for large vocabulary speaker independent recognition systems. It incorporates knowledge of the language, such as its syntactic and semantic information in ASR by providing the probabilities that a word or string of words is/are followed by another word in a given text. The way the words are connected to form sentences is modeled by the language model with the use of a pronunciation dictionary. The language model of our system is a statistical based language model. By assuming that the next word in the sequence depends only upon one previous word, we have created a bigram (2-gram) language model. Finally using this bigram language model, a network which contains words in the training data was created (Young, Gunnar, Mark, Thomas, Dan K., Xunying, Gareth, Julian, Dave, Dan P., Valtchev, Phil 2006).

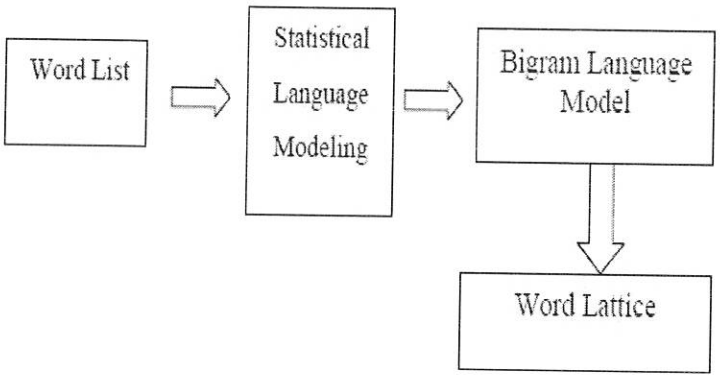


Figure 4.2. Block Diagram of the Language Modeling process (Adapted from (Abraham 2011).

The Language Model used in this study is a backed-off bigram language model. It is developed using the HTK tools HLstats and HBuild. HLstats is primarily used to generate the bigram probability matrix. It reads in a sorted file and generates a bigram language model. The probability matrix is prepared by using the word level transcriptions and statistics on the number of occurrences of each word and each combination of two words. These statistics are then used to create backed-off bigram language models for the training and test using the HBuild tool which translates the gathered statistics into HTK Standard Lattice Format that are used for storing word models and multiple hypotheses from the output of a speech recognizer.

### 4.4.1.4. Acoustic Model

Acoustic models are statistical models which capture the correspondence between a short sequence of acoustic vectors and an elementary unit of speech. The elementary

units of speech that are used in our research are phones. Phones are the minimal units of speech that are part of the sound system of a language, which serve to distinguish one word from another. We use HMM to model the acoustic component in this research. During training, the parameters for the models were estimated from selected records of ARSCo which was taken from Solomon (2005) Amharic speech corpus, which is transcribed at word level. We have created the acoustic model using the audio data of speech and their text scripts and compiling them into a statistical representation of sounds which make up words. This is done through modeling the HMMs. The process of acoustic modeling is shown in Figure 4.3.

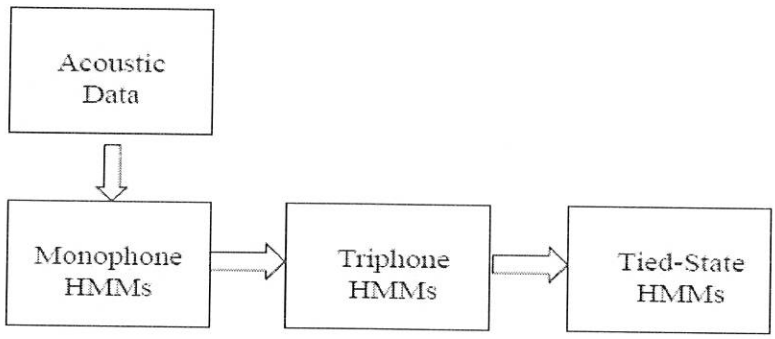


Figure 4.3 Block Diagram of the Acoustic Modeling process (Adapted from (Abraham 2011))

### 4.4.2. Data Preparation

It is plain enough that the required speech data should be prepared and made available before training the models and building the recognizer. Further, the data should pass through different preprocessing tasks in accordance to the requirements

of the various algorithms (tools) in HTK. Thus, the utilized data, the output of the data preparation process, and the preprocessing steps involved are described.

The dataset we used was sample data taken from ARSCo which was prepared by Solomon (2005), which was described in first chapter. Random selection was done and we selected 10 speakers for the training and 5 for the testing, both female and male individuals. We have selected and used total of 471 sentences with 2887 words and 3655 pronunciations for the training and 90 sentences with 847 words and 1022 pronunciations were used for the test. To create the pronunciation dictionary in HTK, we have created a prompts file which is the list of sentences to be recorded, we derive a file from the prompts file which is a sorted list of the unique words that appear in the prompts file and we create the pronunciation dictionary by adding pronunciation information to the words in word list.

The alternative dictionary was done manually by hearing the audio file using the data-driven pronunciation variation modeling approach by taking sample for the purpose of this thesis. We then added pronunciation information (i.e. the phonemes that make up the word) to each of the words in the lexicon, thus creating a pronunciation dictionary. Again the identified word was transcribed manually and added to the dictionary which finally made the alternative dictionary. HTK uses the HDMan command to go through the word lists and look up the pronunciation for each word in a separate lexicon file, and output the result in a pronunciation dictionary. For HTK to compile the acoustic model, we need to make sure that we

have at the very least 3 to 5 usage counts for each phone. If there are phones that only have one occurrence, we must add words that use these phones to the prompts file.

### 4.4.3. Training

Defining the structure and overall form of a set of HMMs is the first step towards building a recognizer. The second step is to estimate the parameters of the HMMs from examples of the data sequences that they are intended to model. The topology for each of the hmm to be trained is built by writing a prototype definition. HTK allows HMMs to be built with any desired topology. HMM definitions can be stored externally as simple text files and hence it is possible to edit them with any convenient text editor. With the exception of the transition probabilities, all of the HMM parameters given in the prototype definition are ignored.

The purpose of the prototype definition is only to specify the overall characteristics and topology of the HMM. The actual parameters will be computed later by the training tools. Sensible values for the transition probabilities must be given but the training process is very insensitive to these. An acceptable and simple strategy for choosing these probabilities is to make all of the transitions out of any state equally likely. In principle the HMM should be tested on a large corpus containing wide range of word pronunciations. For this purpose 471 sentences and 2887 vocabulary with 3655 pronunciations were selected and labeled as it should be formatted for training

data. This was done by defining a prototype model. Since a phoneme based recognizer was built a model represented a phoneme.

The starting point was set of identical monophone HMMs in which every mean and variance is identical. Thus the means and variances of all the states in the model were simply assigned a value of 0 and 1 respectively. The purpose of the prototype definition was only to specify the overall characteristics and topology of the HMM. The actual parameters were computed by the training tools. These were then retrained, short-pause models were added and the silence model was extended slightly. The monophones were then retrained.

### 4.4.4. Recognition and Analysis

HTK provides a recognition tool called HVite that allows recognition using pronunciation dictionary, language models, and output a transcription file against which the recognizer's performance is analyzed. And both the ASR developed using the canonical and the multiple pronunciation (the alternative) dictionaries were tested.

After finishing the recognition the performance of both system with the canonical and alternative dictionary of the Amharic Speech Recognition System pronunciation variation evaluated by running the HResult command which is also another took of HTK and the results found satisfied the objective of this research.

## 4.5. Experimental Results and Analysis

As indicated in the other chapters the starting point of this research was to show the effect of the pronunciation variation on the performance of Automatic Speech Recognition for Ahmharic language.

The results were 60.54% word accuracy and 73.23% words correctness with 13.11% correct sentence for the Canonical dictionary. For the alternative dictionary the results were 74.57% correct words 62.68% accuracy and the same 13.11% correct sentences are found. Here there was no sentence accuracy difference was seen, which basically is because of the limitation of the data used the little variation modeling. Since Amharic has different types of pronunciation variation which are not the focus of this research but still are considered as variation and have an influence in the performance of Speech Recognition of Amharic language. Table 4.1. shows the results of the experiment of the canonical and alternative dictionary.

| | **% of words accurately recognized** | **% sentences correctly recognized** |
|---|---|---|
| Canonical | 60.54 | 13.11 |
| Alternative | 62.68 | 13.11 |

Table 4.1. Table showing the results of canonical and Alternative dictionary

The difference shown in the word accuracy was only 2.14% because of very little phone difference. Therefore, the recognition showed result difference which confirms the correctness of our hypothesis and the effect of pronunciation variations in the development of an Automatic Speech Recognition system for Amharic language. Considering the result and performance of the system we indicated that there are limitations of sampling and the focus of this research is to show the effect of pronunciation variation in Amharic speech recognition using the data-driven approach. Therefore, if possible pronunciation variants were considered there would be better result. Of course our main intension is not to compare or see all variants of the Amharic language.

The other point is simple error analysis was done to identify the causes for the incorrect recognitions for the above two pronunciation dictionaries. Again this is also not the issue of this research but whenever developing an Automatic Speech Recognition Systems checking the result and considering word error rate is one point. When the incorrectly identified utterances were manually compared with the correct ones, on most of the utterances, only few phones happened to be incorrectly identified. This was due to the fact that most of the words in manual utterances miss or add the epenthetic vowel "ኧ"/ix/. And the other point which was not considered here but has its own influence is utterances are incorrectly recognized due to gemination problem.

# CHAPTER FIVE

# CONCLUSION AND RECOMMENDATION

Current ASR systems perform substantially worse for those languages which are very familiar with technology. ASR research has always drawn on knowledge from linguistics, and our motivations is that there are many additional ideas in linguistics to draw on than have been used in Amharic Speech Recognition to date. This paper describes an effect of pronunciation variation on the development of Amharic Speech Recognition Systems. This section discusses the achievements of the research, drawbacks and possible future works to improve the work carried out by this research.

## 6.1. Conclusions

The primary objective of this research was to show the effect of pronunciation variation for Amharic Speech Recognition. In order to meet the research objective sample data was taken from ASRCo corpus of Solomon (2005) and experimental analysis was done. As we are in the early stage of modeling pronunciation variation for Amharic Speech Recognition, we can say that this research achieved the primary goal of this study. The test results indicate that adding words with multiple pronunciations to the lexical using the direct data-driven approach will improve the performance of Amharic Speech Recognition systems. The test results also show that there were 60.54% word accuracy and 73.23% words correctness with 13.11%

correct sentence for the Canonical dictionary and 74.57% correct words 62.68% accuracy and the same 13.11% correct sentences for the alternative dictionary

Here the result for the sentence was the same for both the canonical and alternative dictionaries, which can be taken as the limitation of this research because of very limited pronunciation variables sampling. But still can be a supporting result for the basic objective of this research, because even with very limited addition of pronunciation variants the accuracy and performance of the alternative dictionary showed an improvement. Therefore, if as many pronunciation variants as possible are modeled in Amharic Speech Recognition, there will be observable performance increase and WER reduction.

## 6.2. Recommendations

As indicated in the conclusion part, the performance of a Large Vocabulary Continuous Speaker independent Amharic Speech Recognition System can further be improved by adding pronunciation variants to all words that have pronunciation variations in the pronunciation dictionary. This is due to the fact that the research has got performance increment for the multiple-dictionary after adding alternative pronunciation variants using direct data-driven approach to only some of the word which are identified by the researcher's hearing ability in the pronunciation dictionary.

As mentioned in the discussion part we have used the data-driven approach to add the pronunciation variants and based on the research experimental analysis, recognition results and conclusions, therefore we recommend to transcribe/to include all possible pronunciation variants in the pronunciation dictionary to reduce word error rate in the system and improve the performance of Amharic Speech Recognition system.

We also recommend for researchers to work with the knowledge base lexical modeling approach which needs the knowledge of linguistics to model the pronunciation variants found in Amharic language. As mentioned in chapter 3 Amharic language has different dialects which definitely need to be considered in pronunciation modeling.

Finally, we recommend specially for computational linguists to show and compete both the data-driven and knowledge base approach as a hybrid approach to improve the performance of Amharic Speech Recognition system.

# References


Abdella Kemal. 2010. Speaker Dependent Speech Recognition for Sidaama Language. MSc, Thesis, Department of Information Science, Addis Ababa University, Ethiopia.

Abraham Woubie. 2011. Enhanced Amharic Speech Recognition System. MSc Thesis, Department of Computer Science, Addis Ababa University, Ethiopia.

Adami André G. 2010. Automatic Speech Recognition. *From the Beginning to the Portuguese Language.*, California, USA.

Amanda W. and Rachel W. 2011. Amharic Language and Culture Manual National Language of Ethiopia, *Texas State University*, USA.

Baye Yimam. 2010. አጭርና ቀላል የአማርኛ ሰዋስው. (*A Short and Simple Amharic Grammar).* Addis Ababa: Alpha Printers.

Bender M.L., Bowen, J.D., Cooper and Ferguson, C.A (1976). Language of Ethiopia. *Oxford University press*, UK.

Cremelie N. and Martens J. In search of better pronunciation models for speech

recognition. In Proceedings of the 29th Speech Communication. Trier University, Germany: P.1999.

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

Daniel Jurafsky. 2009. *Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. University of Colorado at Boulder, Upper Saddle River, New Jersey , USA.

Deller J.R. Jr., Hansen, J.H.L. and Proakis, J.G. 2000. Discrete-time Processing of Speech Signals. *Macmillan Publishing Company*, New York, USA.

Derib Ado. 2011. An Acoustic Analysis of Amharic Vowels, Plosives and Ejectives. Ph.D. Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

Huang X., Alleva, F., Hon, H.W., Hwang, M. Y. and R. Rosenfeld.1992. The SPHINX-II speech recognition system. *Pittsburgh.* USA.

Ibrahim Badr. 2011. Pronunciation Learning for Automatic Speech Recognition. Massachusetts Institute of Technology, USA.

Ingunn Adal. 2002. Learning pronunciation variation, A data-driven approach to rule-

based lexicon adaptation for automatic speech recognition. N-7491 Trondheim, Norway.

Ingunn Amda., Eric Fosler. Pronunciation Variation Modeling in Automatic Speech Recognition. *Telektronikk*, Norway: Lussier. 2003.

Javier Ferreiros, Javier Macías-Guarasa, José M. Pardo and Luis Villarrubia. 1998. Introducing Multiple Pronunciations in Spanish Recognition Systems. Universidad Politécnica de Madrid, Spain.

Junqua J.-C, Steven Fincke and Ken Field. "Influence of the Speaking style and the Noise Spectral Tilt on the Lombard reflex and automatic speech recognizers." California, USA: 1993.

Kessens J., Cucchiarini C. and Strik, H. "A data-driven method for modeling pronunciation variation." University of Nijmegen, the Netherlands: 2001.

Kessens J., Wester M. and Strik H. "Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation." *In Proceedings of the 29th Speech Communication. Trier University*, Germany: 1999.

Kinfe Tadesse. 2002. 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.

Markus, Forsberg. 2003. "Why is speech recognition difficult?" Department of Computing Science Chalmers University of Technology, Sweden.

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

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

Murat Sarac_lar., Sanjeev Khudanpur. 2003. Pronunciation change in conversational speech and its implications for automatic speech recognition. Johns Hopkins University, USA.

Peter Ladefoged. 2003. Vowels and consonants. *An introduction to the sounds of languages*. Blackwell publishers inc. 350 Main street, Malden, Massachusetts 02148, USA. Blackwell Publishers Ltd. 108 Cowley Road Oxford OX41JF, UK.

Rune Saetre. 2003. *Natural Language Understanding (NLU)*, N-7491 Trondheim,

Norway.

Sebsibe H/Mariam , S P Kishore, S.P., Alan W Black , Rohit Kumar and Rajeev Sangal. Unit Selection Voice for Amharic Using FESTVOX. *In proceeding of the 5th ISCA Speech Synthesis Workshop, Pittsburgh*, USA: 2004.

Solomon Gizaw. 2006. Multiple Pronunciation Models for Amharic ASR. Msc Thesis, Addis Ababa University, Ethiopia.

Solomon Tefera Abate, Wolfgang, Menzel and Bahiru, Tefla. An Amharic speech corpus for large vocabulary continuous speech recognition. *In Proceedings of the 9th European Conference on Speech Communication and Technology, Interspeech Lisbon,* Portugal: 2005.

Solomon Tefera Abate. 2005. Automatic Speech Recognition for Amharic. Ph.D. Thesis, Hamburg University, Germany.

Solomon T., Martha Y., Wolfgang M. Amharic Speech Recognition: Past, Present and Future. *Proceedings of the 16th International Conference of Ethiopian Studies, ed. by Svein Ege,* Harald: 2009.

Strik H. and Cucchiarini C. "Modeling pronunciation variation for ASR:" A survey of the

literature. University of Nijmegen, The Netherlands: 1999.

Victoriya Fromkin. 2003. *Introduction to Language*. Late, University of California, Los Angeles

Weintraub M., Taussig, K. Hunicke-Smith, and Snodgras, A. Effect of speaking style on LVCSR performance. *In Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, USA: 1996.

Westendorf, C.M. and Jelitto J. J. Learning Pronunciation Dictionary from Speech Data. *In Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, USA: 1996.

Wester M. 2001 Pronunciation modeling for ASR – knowledge-based and data-derived methods. University of Nijmegen, The Netherlands.

Wooters and Stolke 1996. Multiple-Pronunciation Lexical Modeling in a Speaker Independent Speech Understanding. Berkeley, USA.

Yiming Huang. 2009. Phoneme Recognition Using Neural Network and Sequence Learning Model. the Russ College of Engineering and Technology of Ohio University. Ohio.

Young Steve , Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev and Phil Woodland. 2006. *The HTK Book*. Cambridge University Engineering Department, UK.

Zegaye Seifu. 2003. HMM based large vocabulary, speaker independent, continuous Amharic speech recognizer. Msc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia.

Zelalem Getahun. 2007. Amharic Political plays (1974-81) a contextual study. MA. Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

# Appendix A: Sample Alternative Dictionary

baalafawe [baalafawe] b aa l a f a w e

baalafawe [baalafawe] b aa l e f a w e

badabube [badabube] b a d a b u b e

badabube [badabube] b a d e b u b e

bafite [bafite] b a f i t e

bagaaraa [bagaaraa] b a g aa r aa

bagaaraa [bagaaraa] b e g aa r aa

bahagaraacene [bahagaraacene] b a h a g a r aa c e n e

bahagaraacene [bahagaraacene] b aa g a r aa c e n e

bahhadise [bahhadise] b a hh a d i s e

bahhadise [bahhadise] b aa d i s e

bahhaweropaa [bahhaweropaa] b a hh a w e r o p aa

bahhaweropaa [bahhaweropaa] b a hh i r o p e

bahheereteraa [bahheereteraa] b a hh ee r e t e r aa

bahheereteraa [bahheereteraa] b a hh ee r i t e r aa

bahheereteraa [bahheereteraa] b a hh ee r i t e r i y aa

bahhenedihe [bahhenedihe] b a hh e n e d i h e

bahhenedihe [bahhenedihe] b á n e d i h e

bahuaalaa [bahuaalaa] b a h u aa l aa

bahuaalaa [bahuaalaa] b a h u w aa l aa

bakule [bakule] b a k u l e

bamaalate [bamaalate] b a m aa l a t e

bamahone [bamahone] b a m a h o n e

# Appendix B: The Configuration Parameter

TARGETKIND = MFCC_0_D_N_Z

TARGETRATE = 100000.0

SAVECOMPRESSED = T

SAVEWITHCRC = T

WINDOWSIZE = 250000.0

USEHAMMING = T

PREEMCOEF = 0.97

NUMCHANS = 26

CEPLIFTER = 22

NUMCEPS = 12

# Appendix C: Fragment of the tree.hed script

```
RO 100 stats

TR 0

QS "R_NonBoundary" { *+* }
QS "L_NonBoundary" { *-* }
QS "R_Silence" { *+sil }
QS "L_Silence" { sil-*}
QS "R_Stop" { *+b,*+d,*+g,*+p,*+t,*+k,*+pp,*+tt,*+q }
QS "L_Stop" { b-*,d-*,g-*,p-*,t-*,k-*,pp-*,tt-*,q-* }
QS "R_Nasal" { *+m,*+n,*+nn }
QS "L_Nasal" { m-*,n-*,nn-* }
QS "R_Fricative" { *+f,*+s,*+ss,*+h,*+z,*+zz,*+x,*+h }
QS "L_Fricative" { f-*,s-*,ss-*,h-*,z-*,zz-*,x-*,h-* }
QS "L_Affricate" { *+c,*+cc,*+dd }
QS "R_Affricate" { c-*,cc-*, dd-*}
QS "R_Liquid" { *+l,*+r }
QS "L_Liquid" { l-*,r-* }
QS "R_Glide" { *+y,*+w }
QS "L_Glide" { y-*,w-* }
```