Automatic Speech Recognition for Amharic

Dissertationsschrift zur Erlangung des Grades eines
Doktors der Naturwissenschaften
am Fachbereich Informatik
der Universität Hamburg

Vorgelegt von
Solomon Teferra Abate
aus Addis Ababa, Äthiopia

December, 2005

Dedicated

to my Savior Jesus Christ

and

to my beloved wife Marthi (Martha Yifiru Tachbelie)

Acknowledgments

I am truly thankful to my God for every success in my life. He has given me Jesus Christ to save me from my way to hell. He has also given me Marthi (Martha Yifiru Tachbelie), my wife to share the burdens of this earthly life, and Deborah, our daughter to bring us into a state of parenthood.

Prof. Dr. Ing. Wolfgang Menzel merits my heartfelt gratitude for his in-exhaustible help throughout my work on this thesis. He is an accredited and gifted man who has advised, guided, and encouraged me by his critique. I am always praying for his heavenly and earthly blessing.

Prof. Dr. Bairu Tafla and **Prof. Dr. Manfred Kudlek** have also exerted a significant amount of effort in the success of this thesis. They have spent their precious time to read and comment on my thesis. I would like to express my thanks to them. May God bless them in every way.

Martha Yifiru has played the key role in enabling me to use all that God has given me. She is both my partner in prayer and life as a loving wife, mother, sister, and colleague. As a colleague she has given me valuable comments on my research and on writing the draft of my thesis. As a wife, she has tolerated my inability to give her and Deborah their well-deserved quality time and taking on all of the family burdens at home.

I would also like to thank **Daniel Yakob**, who made the archive of the Ethio-Zena website and his SERA to Ethiop converter (namely g2) available to us, and **Berhanu Beyene,**

Manfred Kudlek, Olaf Kummer, and Jochen Metzinger for their Ethiop font. I am also thankful to all people who made the HTK toolkit available online.

The staff of the faculty of Informatics of the Addis Ababa University contributed a lot to the realization of the Amharic corpus. I am grateful for their contribution, especially to **W/ro Ayneabeba Amenu** and **W/ro Kestela Mengistu**. I would also like to extend my gratitude to all readers for the Amharic corpus.

I cannot rush to thank the most important assistant in the work of this thesis without appreciating the helpfulness of NaTS group, especially **Karin Jark, Dirk Knoblauch, Michael Daum, Kilian A. Foth** who helped me in different ways.

Now I can express my most important gratitude to **DAAD (Deutscher Akademischer Austauschdienst)** that provided the financial resources for my work, including all of the costs involved with the speech data collection.

Solomon Teferra Abate

December, 2005

Zusammenfassung

In dieser Arbeit haben wir verschiedene Möglichkeiten zur Entwicklung eines sprecherunabhängigen Spracherkennungssystems mit großem Wortschatz für das Amharische untersucht.

Amharisch ist die offizielle Landessprache in Äthiopien. Innerhalb der Familie der semitischen Sprachen hat es die größte Anzahl von Sprechern nach dem Arabischen. Amharisch ist eine der Sprachen, die eine eigene Verschriftung haben. Es gibt jedoch noch kein Sprachkorpus, das zur Entwicklung eines automatischen Spracherkennungssystems (ASRS) für Amharisch verwendet werden könnte. Daher haben wir ein solches Corpus entwickelt, das für verschiedene Untersuchungen zum gesprochenen Amharisch eingesetzt werden kann.

Unter Verwendung der Sprachsignaldaten haben wir ein ASRS für Amharisch entwickelt, das auf Hidden-Markov-Modellen (HMM) beruht. Diese Arbeiten gingen von der Annahme aus, dass sich die amharischen Silben dank ihrer sehr regelmäßigen CV-Struktur gut als elementare Erkennungseinheiten eignen. Wir waren tatsächlich in der Lage zu zeigen, dass Silbenmodelle als konkurrenzfähige Alternative zu einer Standardarchitektur auf der Grundlage von Triphon-Modellen verwendet werden können.

Die optimale HMM-Struktur für amharische CV-Silben, die wir in unseren Experimenten gefunden haben, ist ein Modell mit 5 emittierenden Zuständen und 12 Gauss'schen Mischverteilungen, jedoch ohne Auslassungen und ohne Sprünge. Unter Verwendung dieser akustischen Modelle, die 15 MByte Speicherkapazität erfordern, wurde zusammen mit dem Sprachmodell und der Sprecheradaption eine Erkennungsgenauigkeit von 90,43% auf den

5.000-Wort-Evaluationsdaten bei einer Geschwindigkeit von 2,4 Minuten pro Satz erreicht.

Unter den getesteten Triphon-Modellen erreichte eine Topologie mit 3 emittierenden Zuständen, Auslassungen und 12 Gauss'schen Mischverteilungen die besten Ergebnisse. Dieses Set von Akustikmodellen erfordert 38 M Speicherkapazität und hat eine Worterkennungsgenauigkeit von 91,31% bei einer Geschwindigkeit von 3,8 Minuten pro Satz.

Wir haben die Ergebnisse unserer Experimente im Hinblick auf Worterkennungsgenauigkeit, Erkennungsgeschwindigkeit und Speicherbedarf analysiert und sind zu dem Schluss gekommen, dass die Modellierung von CV Silben, die den orthographischen Symbolen entsprechen, im Vergleich zu den überwiegend verwendeten phonbasierten Erkennungseinheiten für das Amharische die bessere Alternative darstellt.

Da unsere Arbeit den ersten Versuch zur Entwicklung eines ASRS's für das Amharische darstellt, möchten wir vier Felder nennen, die zukünftig besondere Aufmerksamkeit verdienen: Korpora für gesprochene Sprache, Sprachmodellierung, verbesserte akustische Modelle und die Anwendung von Spracherkennern. Unser Sprachkorpus enthält vorgelesene Sprache und kann daher weder zur Entwicklung eines Erkenners für Spontansprache, noch für telefonbasierte Anwendungen verwendet werden. Die Akustikmodelle unseres Spracherkennungssystems sind zudem wegen Datenmangels nicht ausreichend trainiert. Sowohl die irreguläre Verwendung der Vokale der 6. Ordnung und des Glottisschlags als auch die mögliche Konsonantenlängung werden von unseren Aussprachewörterbüchern nicht behandelt. Aus der reichhaltigen Flektion des Amharischen resultiert auch eine hohe Perplextät unserer Sprachmodelle. Bisher haben wir noch keine Anstrengungen zur Anwendung der von uns entwickelten Erkenner unternommen. Daher empfehlen wir, dass Forscher und Entwickler diesen Gebieten der Spracherkennung für Amharisch ihre verstärkte Aufmerksamkeit zuwenden.

Abstract

In this work we have explored various possibilities for developing a Large Vocabulary Speaker Independent Continuous Speech Recognition System for Amharic.

Amharic is the official language of Ethiopia. Within the Semitic language family, it has the greatest number of speakers after Arabic. Amharic is one of the languages which have their own writing system. There is, however, no speech corpus that can be used for the development of an Automatic Speech Recognition System (ASRS) for Amharic. We, therefore, developed an Amharic speech corpus that can be used for various kinds of investigations into the nature of spoken Amharic.

Using the corpus, we have developed an ASRS for Amharic based on Hidden Markov Models (HMM). The research was guided by the assumption that, due to their highly regular Consonant Vowel (CV) structure, Amharic syllables lend themselves to be used as a basic recognition unit. Indeed, we were able to show that syllable models can be used as a competitive alternative to the standard architecture based on triphone models.

The optimal HMM topology for Amharic CV syllables which we found in our experiments is a model with five emitting states, and twelve Gaussian mixtures without skips and jumps. Using this set of acoustic models, which requires 15MB memory, together with the language model and use of speaker adaptation, we obtained a word recognition accuracy of 90.43% on the evaluation test set at a speed of 2.4 minutes per sentence with 5,000 words.

Among the triphone models tested, a topology with three emitting states, with skips and twelve Gaussian mixtures produced the best results. This set of acoustic models requires 38MB memory and has a word recognition accuracy of 91.31% at a speed of 3.8 minutes per sentence.

We have analyzed the results of our experiments from the point of view of word recognition accuracy, recognition speed and memory requirements, and concluded that for Amharic modeling CV syllables, as represented by the orthographic symbols, is a better alternative to the prevailing modeling units of elementary sounds, like phones.

Since our work is the first attempt in the area of developing ASRSs for Amharic, we would like to mention some of the areas that deserve further investigation: Speech corpora, language model, acoustic models, and the application of the recognizers. Our speech corpus is a read speech corpus and cannot be used to develop, e.g. recognizers for spontaneous speech or telephone-based applications. The acoustic models of our speech recognition system have also suffered from a shortage of training speech data. The irregular realization of the sixth order vowel and the glottal stop consonant, as well as the gemination of the other consonants are not handled by our pronunciation dictionaries. Due to the rich inflection of Amharic, we also have to deal with the relatively high perplexity of our language models. So far, we have not taken any step towards the application of the recognizers that we have developed. We recommend, therefore, that researchers and developers give attention to these areas of speech recognition for Amharic.

Contents

List of Tables

List of Figures

Acronyms

ASR - - Automatic Speech Recognition

ASRS - - Automatic Speech Recognition Systems

CV - - Consonant-Vowel

HMM - - Hidden Markov Modeling

HTK - - The Hidden Markov Model Toolkit

LVSIASRS - Large Vocabulary Speaker Independent ASRS

DARPA - - Defense Advanced Research Projects Agency

Chapter 1

Introduction

Our work is an attempt to contribute to international research and development efforts in the area of Automatic Speech Recognition (ASR), with the general objective of exploring the possibilities for the development of a large vocabulary, speaker independent and continuous speech recognition system for Amharic speech. The hypothesis of our research reads as "*to develop an ASR for Amharic, modeling syllables, as represented by the orthographic symbols, is a better alternative to the prevailing modeling units of elementary sounds, like phones.*"

ASR is one step in the development of an "intelligent machine" that can "listen" to human speech. To automatically listen to human speech, which is termed in literature as Automatic Speech Recognition, a machine is expected to perform two functions: speech recognition (SR) and speech understanding (SU). 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. Since we are not aiming to deal with the wider scope of speech understanding, the definition of Automatic Speech Recognition (ASR) that fits to the scope of our work is "the decoding of the information conveyed by a speech signal and its transcription into a set of characters" (Junqua and Haton 1996).

The development and advancements in computer technology in their storage space, processing speed and other requirements of speech processing have encouraged research and development in ASR. As a result, a considerable amount of research on the development of Automatic Speech Recognition Systems (ASRS) has been conducted and lots of ASRS have been developed. So far, however, speech recognizers have only been developed for a few of the enormous number of human languages and those are still faced with different problems and constraints. Moreover, their performance is poor in noisy environments, and recognition of spontaneously spoken speech is still very difficult. Our research contributes by adding another language, i.e. Amharic, to those covered by ASR research.

1.1 Application and Benefits of Speech Recognition

Companies increasingly use and develop systems with speech recognition interfaces citing various benefits, including (Markowitz 1996):

- increased productivity by enabling a person to use his/her hands and mouth for different tasks and making hands-free work possible,
- rapid return on investments that apply ASRSs to speed up tasks,
- access to new markets (24-hours service),
- environment control (by disabled people, e.g.), and
- the naturalness of communication between man and machine.

Rodman (1999) and other researchers pointed out a number of specific benefits that computer speech recognition is coming up with. Some of them are to:

- enable us to orally dictate our computers;
- automatically translate spoken language;
- enable us to communicate with remote computers (e.g.: Tele-banking, Expert system, Database-query, Information retrieval, etc.);
- have voice-controlled equipment (e.g.: Car, Airplane, Robotics, etc.);
- automatically receive service requests in different organizations, primarily telecommunications;
- assist/aid persons with disabilities;
- support teaching oral skills;
- develop a multimedia Computer-Assisted-Instruction system; and
- analyze speech evidence for the police and the court.

Our research project develops the basis of applying Amharic speech recognition for at least some of the above benefits.

1.2 Automatic Speech Recognition Systems (ASRS)

Speech corpora are the prime source of data for basic and applied research in the area of spoken language communication, and for technology development in the area of Spoken Language Processing (SLP), in general, and ASR, in particular (Schiel and Draxler 2004). The availability of a speech corpus is needed for the development of speech recognition systems. The quality of the corpus has a great impact on the performance of the resulting speech recognizer. That is why great amounts of resources are spent on the preparation of different types of speech corpora in a number of languages. The type of the corpus also determines the type of the resulting speech recognizer. Since there is no Amharic speech corpus, we developed one that can be used for our research project and related basic and applied research works. The details of its preparation are given in Chapter 4. Due to limitations of the required resources and the purpose of a corpus preparation, a number of constraints need to be imposed on a speech corpus and consequently on the resulting speech recognizers. On the basis of the constraints, automatic speech recognition systems may be categorized as follows:

- Speaker dependent or independent;
- Isolated, connected or continuous speech;
- Small, medium or large vocabulary;
- Read or spontaneous speech;
- Noisy or noise free speech.

An ASRS may be speaker dependent, isolated speech, small vocabulary, and be developed to recognize speech that is read in a noise free environment. At this level a number of ASRSs have been developed with satisfactory performance. But higher level systems such as speaker independent, continuous speech, large vocabulary, spontaneous speech, and for noisy speech are still subject of research in this area. ASRSs in different human languages, if they are developed at all, are, therefore, developed under one or more of the constraints mentioned above. The difficulties that enforced these constraints are briefly described in Chapter 3.

1.3 The Amharic Language

Amharic is the official language of Ethiopia. It is a Semitic language that has the greatest number of speakers after Arabic (Hayward and Richard 1999). Amharic has five dialectical variations spoken in five different Amharic speaking regions: Addis Ababa, Gojjam, Gonder, Wollo, and Menz (Cowley, et.al. 1976). The speech of Addis Ababa has emerged as the standard dialect and has wide currency across all Amharic-speaking communities (Hayward and Richard 1999).

Amharic is one of the languages that have their own writing system. The writing system is used across all Amharic dialects. Getachew (1967) stated that the Amharic writing system is phonetic. It allows any one to write Amharic texts if s/he can speak Amharic and have knowledge of the Amharic alphabet. Unlike most known languages, no one needs to learn how to spell Amharic words and to see the word written first in order to know how to spell it. In support of the above point, Leslaw (1995) noted that on the whole, no real problems exist in Amharic orthography, as there is more or less, a one-to-one correspondence between the sounds and the graphic symbols, except for the redundant symbols.

As with all of the other languages, Amharic has its own characterizing phonetic and phonological properties. For example, Amharic has a set of speech sounds that is not found in other languages. Amharic also has its own inventory of speech sounds, some may be common with other languages, that are combined to form its syllables and words. The words themselves are combined to form phrases and sentences of the language according to its syntactic rules. A review of the Amharic phonetics, phonology and writing system, in view of developing ASRS, is given in Chapter 2.

1.4 Achievements of the Research Project

Speech recognition systems have been developed only for technologically favored languages like English, Italian, Japanese, Chinese, and German. Most of the benefits that are offered by speech technology are, therefore, language specific and are not developed and available for speakers of languages, like Amharic. Thus our main achievement is the development of an ASR for Amharic, a language which is not technologically favored.

As a byproduct of our research work on the development of ASR for Amharic, we have prepared an Amharic speech corpus. Therefore, having a corpus that can be used for consecutive research on the development of Amharic speech recognition systems is one of our considerable achievements.

Most important is that our research shows the use of Amharic Consonant-Vowel(CV) syllables as units of speech recognition. It also reflects the capabilities of the HTK[1] speech recognition toolkit with regard to the use of Amharic Consonant-Vowel(CV) syllables as units of speech recognition.

1.5 Scope and Limitation of the Study

Automatic Speech Recognition exceeds the scope of a Ph.D. project. We have, therefore, limited our scope to the area of ASR that transcribes speech into a set of Amharic characters. We have also worked only on a read speech corpus and consequently developed read speech recognizers.

There are also limitations of our work within the indicated scope. We can categorize them into the areas of speech corpus, language model and acoustic models. The details of the limitations in these areas are given in Section 6.3.

1.6 Methodology of the Research

We used the Hidden Markov Modeling (HMM) methods with left-to-right model structure[2] to develop the recognizer. Currently the HMM methods are most popular and dominating methods in speech recognition research. Moreover, most of the current speech recognition systems and well established toolkits, like the HTK toolkit, are using the HMM to represent any unit of speech. Since there are strong temporal constraints in speech, left-to-right HMMs

[1] The Hidden Markov Model Toolkit (HTK) is a portable toolkit for building and manipulating hidden Markov models, especially for speech recognition research.

[2] The HMM and their structure are discussed in Section 3.5

are always used.

When we use HMM to model speech, we need to define the fundamental unit of recognition that a HMM will model. This unit of speech should be not only trainable and well defined, but also relatively insensitive to context. The size of speech unit may range from words to phonemes. Word models are not practical for large vocabulary speech recognition because word models cannot share training data and, therefore, need to be trained individually which requires incredibly large amounts of training speech data. Phone models, on the other hand, suffer from co-articulatory effects. One way to minimize co-articulatory effects is to use larger units of speech. Examples of this include syllables or demi-syllables[3]. A more serious problem with these units is their large number in various languages; for example, English has over 20,000 syllables and over 1000 demi-syllables (Lee 1989). In contrast, Amharic has only 233 distinct CV syllables.

This is the fact behind our hypothesis which we dealt with in comparing the performance of ASRSs that are developed using Amharic CV syllables to those using Amharic phones.

1.7 Organization of the Thesis

This thesis is organized into six Chapters. This Chapter introduces the contents of the other Chapters and includes the objectives, hypothesis and achievements of the thesis as well.

Chapter 2 gives details of the Amharic language with an emphasis to features of the language that are related to the development of ASRSs. It also discusses the linguistic features of Amharic.

The mathematical background of ASR is given in Chapter 3. We have emphasized the areas pertinent to our work. The reasons behind most of our methods and approaches used in the experimental work are given in this Chapter.

[3] Demi-syllables are half a syllable, from the beginning of the syllable to the middle of the vowel, or from the middle of the vowel to the end of the syllable.

Chapter 4 covers the development of the corpus used for this thesis work and a brief description of the corpus is also given. The theories and previous works on the development of corpus are also given in this Chapter.

Our experiments and their results, and the comparison between syllable- and phone-based recognizers are presented in Chapter 5. On the basis of our findings, we have drawn conclusions that are given in Chapter 6 which also forwards some recommendations to the attention of developers and scientists in this area.

Chapter 2

The Amharic Language

2.1 Introduction

This chapter presents a brief introduction to areas of linguistics that are related to the development of ASR, namely phonetics, phonology, morphology and syntax. It also introduces Amharic phonetics, phonology, morphology, syntax and writing system in view of developing ASRS.

Phonetics is the study of speech sounds used in languages 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.

We know from phonetics and acoustics that according to their acoustic features, speech sounds can be categorized into consonants and vowels. These are subcategorized into the smallest units of a language called phones. Phones are one of the most common sub-word recognition units used in the development of automatic speech recognizers.

There are three sub-disciplines of phonetics that study the different 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. Figure 2.1 shows some of these articulators. Articulators are parts of the vocal 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.

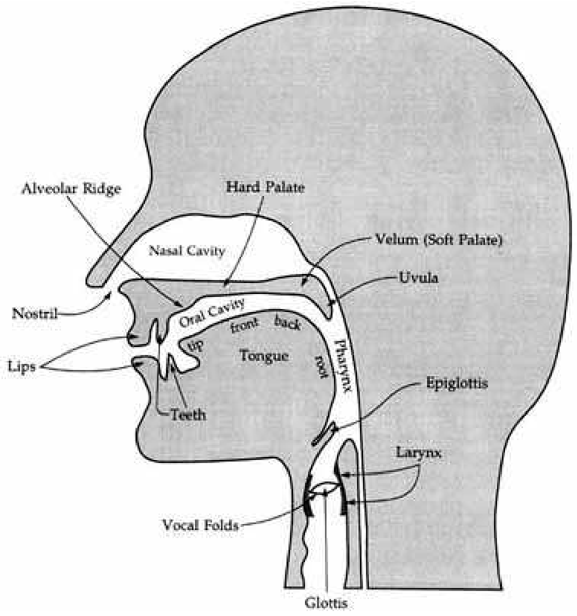

Figure 2.1: Human Vocal Tract

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. Figure 2.2 shows parts of the human ear. An ideal automatic speech recognizer is one 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.

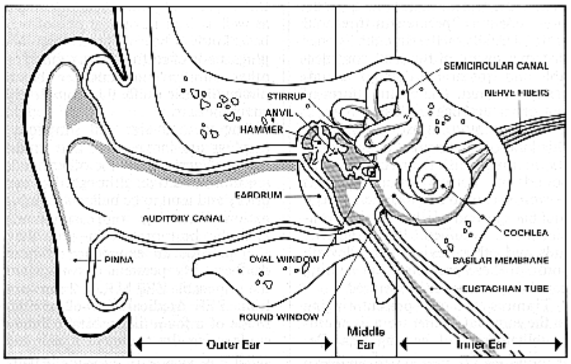

Figure 2.2: Human Ear

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.

Part of the phonological study of a language involves analyzing phonetic transcriptions of speech made by native speakers and trying to deduce what the 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.

Morphology is the study of word formation and structure. It studies how 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 the smallest meaningful unit you can have in a language. The word cats, for example, contains the morphemes cat and the plural -s. In line with the search for an appropriate unit for a language model, morpheme is one option as we can see in different publications, like, Huckvale and Fang (2001) and Vergyri, et.al. (2004).

Syntax is the study of sentence structure. It 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 clearcut 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.

2.2 Amharic Phonetics

Articulatory phonetics shows that characteristics of a sound are determined by the positions of the various articulators in the vocal tract and the state of the vocal cords. Three aspects could be mentioned here: Voicing, Manner and Place of articulation. According to the first aspect we can classify sounds into voiced and unvoiced. In view of the manner of articulation we can have the following classes of sounds: Stops, Fricatives, and Approximants. The places of articulation include Labial, Dental, Palatal, Velar, and Glottal. Vowel sounds are, however, better specified in terms of the following three articulators: Position and height of the tongue, and rounding of the lips.

In accordance to the above understanding of sound classification, a set of thirty eight phones - seven vowels and thirty-one consonants - makes up the complete inventory of sounds for the Amharic language (Baye 1997). The following is a brief overview of each of the major categories of Amharic phones.

2.2.1 Consonants

Amharic has thirty one consonants, which are generally classified as stops, fricatives, nasals, liquids, and semi-vowels (Leslau 2000). Table 2.1[1], page 13 shows the phonetic representation of the consonants of Amharic as to their voicing, place and manner of articulation.

Manner of Articulation	Voicing	Place of Articulation				
		Labials	Dentals	Palatals	Velars	Glottals
Stops	Voiceless	ፕ [p]	ት [t]	ች [tʃ]	ክ [k]	እ [?]
	Voiced	ብ [b]	ድ [d]	ጅ [dʒ]	ግ [g]	
	Glottalized	ጵ [p']	ጥ [t']	ጭ [tʃ']	ቅ [q]	
	Rounded				ኵ $[k^w]$ ጕ $[g^w]$ ቍ $[q^w]$	
Fricatives	Voiceless	ፍ [f]	ስ [s]	ሽ [ʃ]		ህ [h]
	Voiced		ዝ [z]	ዥ [ʒ]		
	Glottalized		ጽ [s']			
	Rounded					ኍ $[h^w]$
Nasals	Voiced	ም [m]	ን [n]	ኝ [ɲ]		
Liquids	Voiced		ል [l] ር [r]			
Semi vowels	Voiced	ው [w]			ይ [j]	

Table 2.1: Categories of Amharic Consonants

[1] In this Table and in Table 2.2, page 15, symbols in square brackets are IPA representations of Amharic phones, using an ASCI equivalent for some IPA symbols: S for ʃ, Z for ʒ, N for ɲ, I for ɨ and A for ə.

According to Leslaw (2000), the Amharic consonants whose phonetic transcription is similar with that of the English consonants are: ብ [b], ድ [d], ፍ [f], ግ [g], ህ [h], ክ [k], ል [l], ም [m], ን [n], ፕ [p], ር [r], ስ [s], ት [t], ቭ [v], ው [w], ይ [j], and ዝ [z]. They correspond to the following English consonants, respectively: b, d, f, g (as in 'gold'), h, k, l, m, n, p, r, s, t, v, w, y, and z. In addition to those, there are sounds that are the same or nearly the same as the English sounds, but are rendered for Amharic by special IPA symbols. These are:
ች [tS] corresponds to ch in 'church';
ጅ [dZ] corresponds to j in 'joke';
ኝ [N] corresponds to ni in 'onion';
ሽ [S] corresponds to sh in 'shoe';
ዥ [Z] corresponds to s in 'pleasure'.

Amharic ድ [d] and ት [t] are pronounced somewhat differently from that of the English d and t. In Amharic they are of the dental type. The glottal stop እ [?], which is the vowel carrier of the language, corresponds in pronunciation to the sound of English 'uh-uh' used as a negation or as 'oh-oh' as an expression of surprise and warning. In Amharic it occurs only medially between vowels in some words such as ብእር [bI?IrI] 'pen' and ሰአት [sA?atI] 'hour'.

The existence of glottal, palatal, and labialized consonants is identified as distinguishing feature of the Ethiopian languages, including Amharic, from foreign languages (Cowley et al. 1976).

The sounds that are characteristic of Amharic and are not found in English are: ጵ [p'], ጥ [t'], ጽ [s'], ጭ [tS'], and ቅ [q] (Leslaw 2000). These sounds are glottalized. To produce these sounds, the stream of air coming from the lungs is shut off by the closure of the glottis, and the air above is then forced out through a stricture formed in the vocal organ. The stricture is at the lips for ጵ [p'], at the teeth for ጥ [t'] and ጽ [s'], at the palate for ጭ [tS'] and at the velum for ቅ [q]. They have a sharp, click-like character. These five glottal consonants correspond to the ordinary or plain consonants ፕ [p], ት [t], ስ [s], ች [tS], and ክ [k] respectively.

The existence of palatal consonants can be considered as another phonetic feature of Amharic. There are six palatal consonants in Amharic - ች [tS], ጅ [dZ], ጭ [tS'], ሽ [S], ዥ [Z] and ኝ [N] which contrast with the corresponding dental consonants ት [t], ድ [d], ጥ [t'], ስ [s], ዝ [z], and ን [n] respectively.

Furthermore, Amharic has four labialized consonants that are pronounced with a slight rounding of the lips. These are ጕ [g^w], ኵ [k^w], ቍ [q^w], and ኍ [h^w]. Although they are used only with the vowel ኣ [a], nearly all the other consonants can be pronounced with a slight rounding of lips.

The consonants ቭ [v] and ፕ [p] appear only in modern loan-words. For example ቪዛ [viza] from visa and ፖሊስ [polis] from police. The consonant ጵ [p'] appears in Greek loan-words through Geez[2]. For example ጠረጴዛ [t'ArAp'eza] 'table'.

All the consonants, except እ [?] and ህ [h], can occur either in geminated or non-geminated form. Gemination is most conveniently described as lengthening in time of the consonant that varies from a slight lengthening to much more than doubling.

2.2.2 Vowels

Amharic has 7 vowels: ኧ [A], ኡ [u], ኢ [i], ኣ [a], ኤ [e], እ [I], and ኦ [o]. Their category is given in Table 2.2, page 15.

	front	center	back
high	ኢ [i]	እ [ɨ]	ኡ [u]
mid	ኤ [e]	ኧ [ə]	ኦ [o]
low		ኣ [a]	

Table 2.2: Categories of Amharic Vowels

From these vowels, ኢ [i], እ [I] and ኡ [u] are high vowels. ኣ [a] is a low vowel and ኤ [e],

[2] Geez had been the spoken language until the end of the Aksum Empire in the ninth century. It was replaced by several new languages (especially by Tigrigna in the north and Amharic in the south). Today, Geez is still used in the Orthodox Church as the language of worship and sacred literature.

ኧ [A] and ኦ [o] are mid vowels. The Amharic vowels ኡ [u] and ኦ [o] are rounded, while the others ኣ [a], ኤ [e], እ [I], ኢ [i] and ኧ [A] are unrounded.

According to Leslaw (1995), the vowels of Amharic are 'pure' vowels in that they do not have the off-glides sometimes characteristic of English vowels like /ey /('long /a/') /ow/ ('long /o/') or /iy/ ('long /e/'). The Amharic vowels /እ/ [I] and /ኦ/ [o], however, often have a kind of on-glide in the pronunciation of some speakers.

Leslaw (2000) noted that there is no precise correspondence in the pronunciation of the Amharic and English vowels. However, he gave the following rough equivalence:

1. The vowel /ኧ/ [A] is pronounced like /e/ in 'bigger': The word ነገ [nAgA] 'tomorrow' can be transcribed as ንኧግኧ [nI?AgI?A]. Similarly, ገረድ [gArArAdI] 'servant' can be transcribed as ግኧርኧድ [gI?ArI?AdI?A]. No word in Amharic begins with initial ኧ [A] except ኧረ [?ArA] 'why, so then'. This vowel has a phonetic variant that tends toward /o/ when preceded by the labial /w/ or by a labiovelar consonant ending in ኧ [A], such as ቈ [q^w]; thus, ወንድም [wAnIdImI] 'brother' may be written and pronounced as ዎንድም [wonIdImI]. They are transcribed as ውኧንድም [wI?AnIdImI] and ውኦንድም [wI?onIdImI], respectively.

2. The vowel ኡ [u] is pronounced approximately like the English /o/ in 'who'. For example, ሁለት [hulAtI] 'two';

3. The vowel ኢ [i] is pronounced like the /ee/ in English 'beet', but without the /y/ glide of the English. For example, ፊት [fitI] 'face';

4. The vowel ኣ [a] is pronounced approximately like the vowel /a/ in the English 'father'. For example, ማታ [mata] 'night';

5. The vowel ኤ [e] has a pronunciation approximately like that of the vowel /a/ in the English 'state', but the consonant preceding this vowel may be frontal so that the consonant has a glide /y/. For example, ቤት [betI] 'house';

6. The vowel እ [I] is pronounced approximately like the /e/ in the English ’roses’. For example, ምላስ [mIlasI] ’tongue’;
7. The vowel ኦ [o] is pronounced approximately like the /o/ in the English ’nor’, but the consonant preceding this vowel may be slightly labialized or rounded so that the consonant has a slight /w/ glide. For example, ሞላ [mola] ’filled’.

If two vowels meet, various changes take place:

1. One of the vowels can be elided. For example, የም [jAmI] plus እሰብር [?IsAbIrI] becomes የምሰብር [jAmIsAbIrI]
2. Depending on the nature of the vowels, a glide ው [w] or ይ [j] appears. For example:
   (a) ውሻ [wISa] - ’dog’ plus ኦች [?otSI] ’suffix for plural’ becomes ውሻዎች [wISAwotSI] - ’dogs’.
3. A glottal stop may appear as follows ብእር [bI?IrI] - ’pen’; ሰአት [sA?atI] - ’hour’; መአት [mA?atI] - ’too much/many’

2.3 Amharic Phonology

All Amharic consonants and vowels are phonemes in that when a consonant or a vowel is replaced with another consonant or vowel within the same environment the change brings another meaning.

ሰባት [sAbatI] - ’seven’ and ሰኣት [sA?atI] - ’hour’

በሳል [bAsalI] - ’matured/ripe’ and በአል [bA?alI] - ’a holiday’

ጐደለ [g^wAdAlA] - ’diminish’ and ገደለ [gAdAlA] - ’kill’

ኰሰሰ [k^wAsAsA] - ’become meager’ and ከሰሰ [kAsAsA] - ’accuse’

ቈጠረ [q^wAt‘ArA] - ’count’ and ቀጠረ [qAt‘ArA] - ’hire’

ስስት [sIsItI] - 'excessive desire' and ሶስት [sosItI] - 'three'
እራት [?IratI] - 'diner' and እሬት [?IretI] - 'bitter'

M. Cohen, N.V.Yushmanov and E. Ullendorff (1965) regard the four labialized consonants as phonemes only if they precede the vowel አ [a]. They point out that in modern pronunciation these consonants, having lost the element of labialization before the vowels ኧ [A], ኢ [i], ኤ [e], እ [I], turned into homogeneous combinations of unlabialized consonant + vowel.

Leslaw (2000) also points out that a labiovelar in any position followed by ኧ [A] may become a plain velar followed by the labial round vowel ኦ [o] without any change in the meaning of the word. Thus ቈጠረ [q^wAt'ArA] becomes ቆጠረ [qot'ArA] - 'counted'; ጐረፈ [g^wArAfA] becomes ጎረፈ [gorAfA] - 'flooded'; and ኰነነ [k^wAnAnA] becomes ኮነነ [konAnA] - 'condemn'. Similarly, a labiovelar in any position followed by እ [I] usually becomes a plain velar followed by the labial rounded vowel ኡ [u]. For example, ቍርስ [q^wIrIsI] becomes ቁርስ [qurIsI] - 'breakfast'.

In Amharic gemination is also phonemic. The following examples give minimal pairs of the geminated, which is represented with doubling the consonant, and non-geminated consonant ን [n], respectively:
ዋና [wanna] - 'chief' as compared to ዋና [wana] - 'swimming';
ገና [gAnna] - 'Christmas' as compared to ገና [gAna] - 'still yet'.

Gemination is at times optional. For example አትሰብርም [?AtIsAbIrImI] - you will not break - the syllable ት [tI] can be geminated or non-geminated.

When two identical consonants are in contact, that is, with no vowel between them, only one consonant is written, but it is pronounced with gemination. For example ምን [mInI] ነው [nAwI] - what is it? - is written as ምነው [mInAwI].

There are phonological variations between the five dialects of Amharic. Hayward and Richard (1999) noted that the most divergent dialect is that of Gojjam province, though the Menz and Wollo varieties also show their own marked features, especially in phonology.

For example:

1. the አ [a] vs ኤ [e] contrast is neutralized after palatals. As ቤት [betI] in Addis Ababa and Gonder vs ብየት [bIjAtI] in others - 'house',

2. the general palatalizing of ድ [d] before ኧ [A] and replacing ኧ [A] by ኤ [e] is characteristic of Menz and Wollo. Illustration: ደበደበ [dAbAdAbA] in others vs ድየበድየበ [dIjAbAdIjAbA] - 'he beat up',

3. in the dialects of Menz, generally, ቀ [q] is replaced by እ [I] and ከ [k] by ሀ [h] whenever these stops occur non-geminated and in non-initial positions. As ለቀን [lAqAnI] in others ለኧን [lA?Ane] in Menz - 'for a day'; and ነከሶ [nAkIso] in others ነሀሶ [nAhIso] in Menz - 'he having bitten', and

4. in Gojjam ብ [b] gives way to ው [w] in positions following a vowel and preceding a consonant. As ገብቷል [gAbIt^wallI] in others ገውቷል [gAwIt^wallI] in Gojjam - 'he has entered'.

2.4 Amharic Morphology

Especially in its verbs, Amharic uses the root-pattern[3] morphological phenomenon. The words in the language consist of stems and affixes. Note that the stem forms can take on various affixes.

For example:

1. the subject is marked on the verb using subject suffix pronouns, as in ሰበርኩ [sAbArIku] - 'I broke'; ሰበርን [sAbArInI] - 'we broke'; ሰበርክ [sAbArIkI] - 'you(second person masculine) broke'; ሰበርሽ [sAbArISI] - 'you(second person feminine) broke'; ሰበሩ [sAbAru] - 'you(second person respected) broke'; ሰበራችሁ [sAbAratSIhu] - 'you(second person

[3] A root is a set of consonants. A pattern consists of a set of vowels which are inserted among the consonants of the root.

plural) broke'; ሰበረ [sAbArA] - 'he broke'; ሰበረች [sAbArAtSI] - 'she broke'; ሰበሩ [sAbAru] - 'they broke',

2. the direct object is marked on the verb, as in ሰበረኝ [sAbArANI] - 'he broke me'; ሰበረን [sAbArAnI] - 'he broke us'; ሰበረህ [sAbArAhI] - 'he broke you(second person masculine)'; ሰበረሽ [sAbArASI] - 'he broke you(second person feminine)'; ሰበሮት [sAbArotI] - 'he broke you(second person respected)'; ሰበራችሁ [sAbAratSIhu] - 'he broke you(second person plural)'; ሰበረው [sAbArAwI] - 'he broke him'; ሰበራት [sAbAratI] - 'he broke it'; ሰበራቸው [sAbAratSAwI] - 'he broke them',

3. some prepositional phrase complements are optionally marked on the verb as in እስክትመጣ ጠብቃለሁ [?IsIkItImAt'a t'AbIqalAhu] - 'I wait until she comes' or 'I wait until you (second person masculine) come'.

4. functional elements like negation maker, conjunction and some auxiliary verbs are also bound morphemes that are attached to the verb:

   - አልሰበርም [?allIsAbArImI] - 'I will not be broken'
   - ሲሰብር [sisAbIrI] - 'while he was breaking'
   - ተሰብሬአለሁ [tAsAbIre?alAhu] - 'I have been broken'

Verb arguments may be indicated by suffix pronouns and a word may stand alone as a sentence. Illustration: ተሰብሯል [tAsAbIr^wallI] - 'It is broken'

The definite article in Amharic is a bound morpheme which is attached to a noun or to the first inflected element in a noun phrase. Illustration: የተሰበረው [jAtAsAbArAwI] - 'The one that is broken'

There are morphological variations between the five Amharic dialects. Illustration: ሲሰብር [sisAbIrI] in Addis Ababa and Gonder becomes ቲሰብር [tisAbIrI] in others - 'while he was breaking'.

We can see from the morphological features of Amharic, as described here, and in literature such as (Titov 1976), that it is one of the highly inflected languages. As it will be discussed in Chapter 5, such property of a language has an impact on the development of ASRS because it increases the size of the pronunciation dictionary and the rate of 'out-of-vocabulary words', and also the perplexity of the language model.

2.5 Amharic Syntax

In the main and subordinate clauses of Amharic the normal word order is SOV, S = Subject, O = direct Object and V = Verb. This is unlike English which is SVO or VSO in Classical Arabic. Modifiers generally precede the words they modify. For example, the difference between Amharic and English, in this aspect, can be seen in the construction of relative clauses such as: ሰለሞን ደብዳቤውን ጻፈ [sAlAmonI dAbIdabewIn s'afA] which means, 'Solomon wrote the letter'. Its word-to-word translation is "Solomon the letter wrote". The other characteristics of Amharic syntax, as mentioned by Cowley et.al (1976), are:

1. Postpositions: Amharic has a set of postpositions like, ውስጥ [wIsIt'I] 'inside', ላይ [lajI] 'on', ጋር [garI] 'with', አጠገብ [?at'AgAbI] 'near'. Nouns followed by these postpositions usually have at the same time the prefix በ [bA], which bears the meaning of in, at, by or the prefix ከ [kA] which means from. For example: ከጠረጴዛው ላይ ውሰድ [kAt'ArAp'ezawI lajI wIsAdI] - 'take from the table'; ጠረጴዛው ላይ አስቀምጠው [t'ArAp'ezawI lajI ?asIqAmIt'AwI] - 'put it on the table'.

2. Limited use of the plural in nouns: Although Amharic has a morphological category of plural in nouns, the plural form is not usually used with numerals or words indicating quantity. In languages such as English, if a numeral other than 'one' is used with a noun, then the noun must be in the plural, e.g. five houses. Amharic allows phrases like five house - አምስት ቤት [?amIsItI betI], many house - ብዙ ቤት [bIzu betI] to mean five houses and many houses, respectively. Note that the plural form of ቤት [betI] is ቤቶች [betotSI].

3. There is a tendency for verbs in subordinate clauses to lack the full specification of

tense, person, and so on, to be made explicit in the main clause. The Amharic verb has a system of five tenses used in the main clauses:

(a) Present-future - ይሰብራል [jIsAbIralI] 'he breaks, is breaking, will break'

(b) Past - ሰበረ [sAbArA] 'he broke'

(c) Past Continuous - ይሰብር ነበር [jIsAbIrI nAbArI] 'he was breaking'

(d) Perfect - ሰብሯል [sAbIr^walI] 'he has broken'

(e) Past Perfect - ሰብሮ ነበር [sAbIro nAbArI] 'he had broken'

Normally, in subordinate clauses the system is reduced so that only tenses (a) and (b) are used and tense (a) has a special 'short' form.

2.6 Amharic writing system

Getachew (1967) stated that the Amharic writing system is phonetic, therefore, it allows anyone to write Amharic texts so long as the language can be spoken and that a good working knowledge of the Ethiopian script is in place. Unlike most known languages, no one needs to learn how to spell Amharic words, nor to see a word first written in order to know how to spell it. In support of this, Leslaw (1995) noted that on the whole, no real problems exist in Amharic orthography, as there is more or less a one-to-one correspondence between the sounds and the graphic symbols, except the redundant ones.

Many (Bender 1976; Cowley 1976; Baye 1986) have claimed the Amharic orthography as a syllabary for a relatively long period of time. Recently, however, Taddesse (1994) and Baye (1997), who apparently modified his view, have argued it is not. Both of these arguments are based on the special feature of the orthography; the possibility of representing speech using either isolated phoneme symbols or concatenated symbols.

In the concatenated feature, commonly known to most of the population, each orthographic symbol represents a detached consonant and a vowel, except for the sixth order[4], which is sometimes realized as a consonant without a vowel and at other times a consonant with a vowel. This representation of concatenated speech sounds by a single symbol has been the basis for the claim made of the writing system, as syllabary. However, the Amharic writing system also has the potential to represent these concatenated symbols using phonemic symbols as stated by Baye (1997).

As anyone who uses the language can easily observe, people are more aware of the concatenated representation than they are of the segments, the individual consonants, and the vowels. They can tell very easily how many CV syllables (ፊደል) [fidAlI] are there in a word. However, if we ask people to count the number of individual sounds heard they often cannot do it. Research made by Baye (1997) ascertained this fact.

The Amharic orthography, as represented in the Amharic Character set - also called ፊደል [fidAlI] - consists of 276 (231 + 20 + 18 + 7) distinct symbols. In addition, there are twenty numerals and eight punctuation marks. These symbols could be classified into four. In the first category there are thirty-three core orthographic symbols, each of which has seven different shapes, usually known as orders, to represent the seven vowels. Each consonant and the seven vowels in combination represent syllables. In this category there are 231 syllables (33 × 7). Four symbols make up the second category for the labio-velars, consisting of five orders and numbering twenty. The eighteen labialized consonants are the third category in the orthography. The grapheme /ቨ/ with its expanded seven orders is the fourth category.

Research in speech recognition should only consider distinct sounds instead of the orthographic symbols, unless there is a need to develop a dictation machine that includes all of the orthographic symbols. Therefore, redundant orthographic symbols that represent the same syllabic sounds can be eliminated. In Amharic there are four graphemes (ሀ, ሐ, ኀ

[4] An order in Amharic writing system is a combination of a consonant with a vowel represented by a symbol. A consonant has therefore, 7 orders or different symbols that represent its combination with 7 Amharic vowels. The order of the vowels is ኧ [A], ኡ [u], ኢ [i], ኣ [a], ኤ [e], እ [I], and ኦ [o]

and ኸ) representing the h (as in he) sound, two graphemes (ሠ and ሰ) that represent the s (as in speech) sound, two graphemes (አ and ዐ) that depict the a (as in an) sound and two graphemes (ጸ and ፀ) depicting the ts (which has no equivalent sound in English) sound. Thus, eliminating these redundant graphemes with their seven orders (6 × 7 = 42), we are left with 234 graphemes. Again, the 1st and the 4th order graphemes of ሀ and አ represent the same sound, therefore, only the first one is needed. The first order (ኸ) of the ኸ sound is rather distinct and should be considered. Hence, there remains a total of 233 distinct CV syllable characters.

Chapter 3

Techniques of Automatic Speech Recognition

3.1 Introduction

As it has been defined in chapter one, the aim of automatic speech recognition (ASR) is to transform a given spoken utterance into the corresponding transcription. Figure 3.1, page 26 shows a general overview of an ASRS. In the figure, AM=the Acoustic Model; LeM=the Lexical Model, which is the pronunciation dictionary; and LaM=the Language Model.

Before an ASRS can be used, it has to learn the characteristics of speech patterns from a speech corpus. That requires the work of training/development of the recognizers. The development of a Large Vocabulary Speaker Independent ASRS (LVSIASRS) involves the development of the acoustic, lexical and language models using a proper speech corpus (large speech database with accompanying transcriptions) and text corpus.

However, before speech data can be used in training or recognition, it must be converted into the appropriate parametric form. The speech parameterization block is used to extract the relevant information from the speech wave form that can be used for discriminating

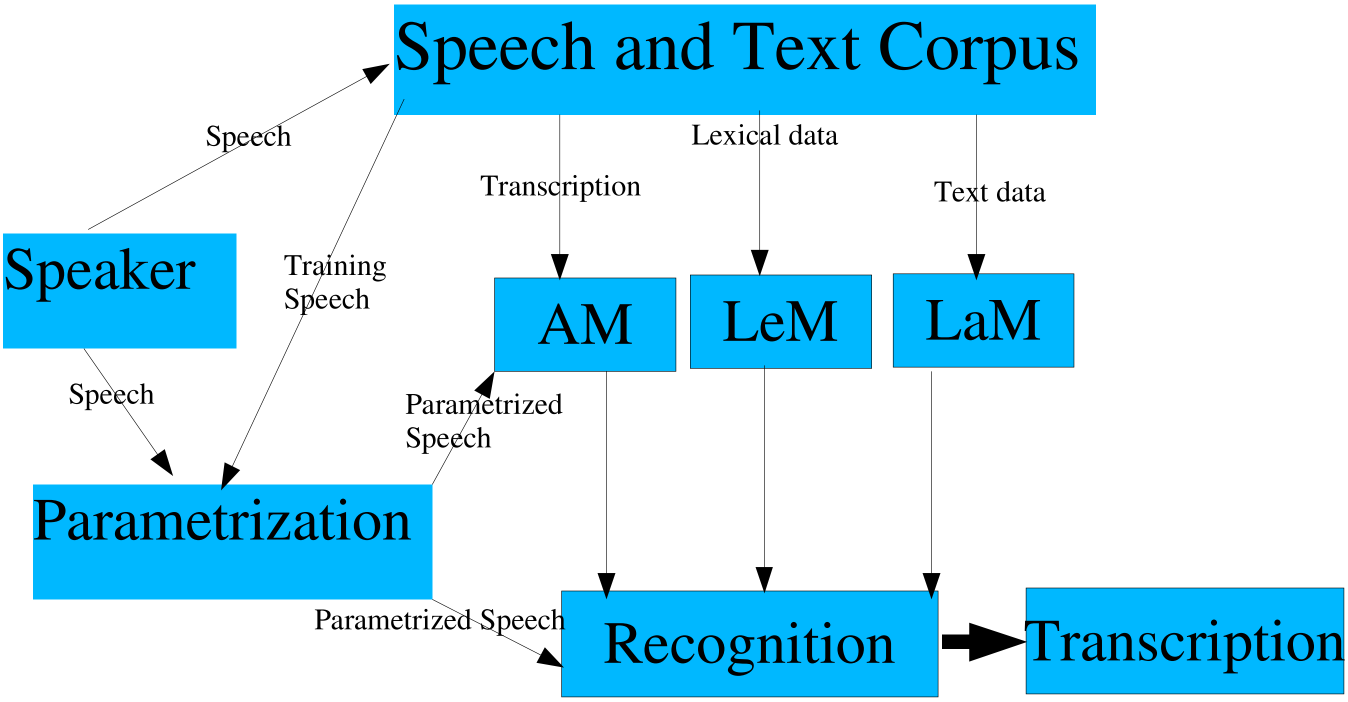

Figure 3.1: ASRS General Structure

among different speech sounds. The information is presented as a sequence of parameter vectors. A part of this chapter presents some of the most common techniques that are applied to speech parameterization/representation.

We have noted in Section 1.2 that according to the constraints imposed during the development of the acoustic, lexical and language models, there are different types of speech recognition systems. This chapter briefly describes some of them. For details, readers are referred to standard reference books such as Rabiner and Juang (1993), Lee (1989), Junqua and Haton (1996), Deller, Proakis and Hansen (1993).

During the use of a LVSIASRS for speech recognition, the acoustic models combined with the lexical and language models are used to determine the most likely transcription of speech. To develop a well-performing ASRS, diverse approaches have been explored. In this Chapter, we give a brief description of the acoustic-phonetic, the artificial intelligence and the pattern recognition approach.

A set of acoustic models (hidden Markov models (HMMs), in our case) is trained, each corresponding to one speech unit (recognition unit). In order to build a set of HMMs, a set of speech data files and their associated transcriptions are required. Any associated transcription must also have the correct format and use the required sub-word or word labels. In addition, a lexical model is prepared to describe how the words are built up from the basic speech units, as well as, a language model describing the sequential relationship between words.

The language model incorporates syntactic and pragmatic knowledge of the language in the ASRS. It is essential to have a satisfactory performance of large vocabulary speech recognition systems. Various procedures can be applied in the development of different types of language models to achieve the objectives that arise. Since statistical language models are most commonly used and adopted for this research, a description is also given herein.

3.2 Parameterization of Speech

The analog form of speech cannot be processed directly by a speech recognizer. The spoken input which is converted into an electrical signal by a microphone, has to be converted into digital form. This analog-to-digital conversion includes sampling, quantization, and coding processes. It is achieved by using analog-to-digital converters (Markowitz, 1996). The digital wave form is analyzed by the use of its spectral features. These spectral features are extracted using different spectral analysis methods. The number of features to be extracted depends on the purpose of its use. Most speech recognition systems, including HMM-based ones, rely on relatively simple speech features.

The purpose of feature extraction is reduction of speech data size. This is done by capturing only the non-redundant parameters of the speech signal as well as obtaining the invariant (with respect to speaker and channel), and dynamic (with respect to time) parameters of the speech signal required for the classification of any unknown speech signal by a classifier, like HMM. The standard way of feature extractions consist of the following steps (C'ernocky'2002):

1. Segmentation: The speech signal is divided into segments where the wave form can be regarded as stationary. The classifiers generally assume that their input is a sequence of discrete parameter vectors where each parameter vector represents just one such segment - frame.

2. Spectrum: Current methods for feature extractions are mostly based on the short term Fourier spectrum and its changes over time, therefore, the power or magnitude Fourier spectrum is computed in the next step for every speech segment.

3. Auditory-like modifications: Modifications inspired by physiological and psychological findings concerning human perception of loudness and sensitivity in relation to frequencies, are performed on the spectrum of each speech frame.

4. Derivatives: Feature vectors are usually completed by the first and second order derivatives of their time trajectories (Delta and acceleration co-efficients). The Delta coefficients describe the change of the feature vectors, while the acceleration co-efficients portray the speed of changes.

There are three basic classes of techniques used to extract speech parameters that are suitable for ASRS. These are Fourier transformation, filter bank analysis and linear predictive coding (LPC).

3.2.1 Fourier Transforms

The Fourier transform provides a representation in terms of amplitude and phase as functions of the frequency parameter f. As the speech signal is non-stationary, a short-time Fourier transform is used.

Signals are converted from time or spatial domain to the frequency domain usually through the Fourier transform. The most common purpose for analysis of signals in the frequency domain is the analysis of signal properties. In Fourier transform, the signal information is converted to a magnitude and phase component of each frequency. Fourier transformation is generally defined by the following equation:

$$S_n(e^{j\varpi}) = \sum_m s(m)e^{-j\varpi m}w(n-m), \tag{3.1}$$

where $\varpi = 2\pi f$, and $s(m)w(n-m)$ is a windowed version of the signal.

Regularly, the Fourier transform is converted to the power spectrum which is the magnitude of each frequency component squared.

3.2.2 Filter Bank

The filter bank passes the digitized speech signal through a bank of bandpass filters whose coverage spans the frequency range of interest in the signal. The filter bank separates the signal frequency bandwidth into a number of frequency bands where the signal energy is measured. Rabiner and Juang (1993) states that since the purpose of the filter-bank analyzer is to measure the energy of the speech signal in a given frequency band, each of the bandpass signals is passed through a non-linearity. The non-linearity shifts the bandpass signal spectrum to the low-frequency band as well as creating high-frequency images. A low pass filter is used to eliminate the high-frequency images, giving a set of signals which represent an estimate of the speech signal energy in each of the frequency bands. To implement the filter-bank, as applied in HTK, the window of speech data is transformed using a Fourier transform and the magnitude or the power is taken (Young et.al. 2002).

3.2.3 Linear Predictive Coding (LPC)

The basic idea behind the LPC model is that a given speech sample at time $n, s(n)$, can be approximated as a linear combination of the past p speech samples, such that

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p(n-p), \tag{3.2}$$

where the co-efficients $a_1, a_2, \cdots, a_p$ are assumed to be constant over the speech analysis frame.

Rabiner and Juang (1993) stated that LPC is one of the most powerful and dominant speech coding techniques. It provides a reliable, accurate and robust method of estimating parameters that characterize speech signal. The importance of LPC lies in its ability to provide extremely accurate estimates of speech parameters, speed of computation and low storage requirement. Markowitz (1996) also indicated, the fact, that these advantages of LPC helped it to serve as the basis for other forms of coding such as cepstral co-efficients and vector quantization.

3.3 Types of ASR

Speech recognition systems can be classified on the basis of the constraints under which they are developed and which they consequently impose on their users. These constraints include: speaker dependence, type of utterance, size of the vocabulary, linguistic constraints, type of speech and environment of use.

One or more of these constraints can be put on a speech recognizer. Ideally, a speech recognition system should be free from any constraint. For example, a speech recognizer can be speaker independent, continuous speech, very large vocabulary and spontaneous speech. Some of the aforementioned constraints and the problems that enforce them are presented briefly. Before we present these constraints, let us give the time frame of development from constrained ASR to the unconstrained ASR.

- The first ASR that has been developed, in 1952, at Bell Laboratories, was able to recognize only isolated digits spoken by a single speaker.
- An ASR that is free of these constraints (speaker dependence and type of utterance), i.e. a system that recognizes continuous speech without limiting the number of speakers, was developed in the 1980s, in the community of the Defense Advanced Research Projects Agency (DARPA).
- In the 1990s, an increasing interest for speech processing under noisy or adverse conditions and for spontaneous speech recognition emerged.
- The recent direction of research in speech recognition is its integration into dialogue systems, including communicating over the telephone.

Speaker dependence: A speaker-dependent speech recognition system requires the users to be involved in its development. On the other hand, speaker independent systems can be used by anybody. However, the latter systems usually perform much worse than the former when they are used to recognize speech from a person who was not involved in their development. This is due to the fact that the acoustic variation between different speakers is very difficult to describe and to model. Lee (1989), pointed out three approaches to make a system speaker independent. The first one is the use of knowledge engineering techniques to find perceptually motivated speech parameters that are relatively invariant between speakers. The second is the use of multiple representations for each reference to capture the between-speaker variations. The final approach is the speaker adaptation, which is adopted in this work.

Speaker adaptation is the modification of model parameters using a small number of adaptation sentences from the new speaker so that the parameters are adjusted to the new speaker. Speaker adaptation techniques examine the speech from the new speaker, and determine the differences between his way of speaking and the 'average' way of speaking, which is reflected by the speaker-independent models. Once these differences are known, either the speaker-independent models or the incoming features are modified, such that they better match the new speaker's acoustics.

If a large amount of adaptation data is available from the new speaker in advance, batch adaptation can be performed. In this kind of adaptation, all data is processed in one step. If no speech data is available beforehand, then in-coming data has to be used while the speaker is using the system, to adapt the models after each (set of) utterance(s) with a very small amount of adaptation data in an incremental way. The adapted models are then used to recognize the following utterance. This method is referred to as incremental, or online adaptation.

Type of utterance: A speech recognizer may recognize every word independently. It may require its user to speak each word in a sentence separating them by artificial pause, or it may allow the user to speak in a natural way. The first type of system is categorized as an isolated word recognition system. It is the simplest form of a recognition strategy, but requires a cooperative speaker. It can be developed using word-based acoustic models without any language model. If, however, the vocabulary increases and sentences composed of isolated words are to be recognized, the use of sub-word acoustic models and language models become important.

The second type of systems is a connected speech recognition system. Deller, Proakis and Hansen (1993) noted that the term 'connected speech' refers to the recognition strategy rather than to the speech itself, because the speech is uttered in a continuous manner. The performance of a connected speech recognition system, however, can be affected by the cooperativeness of the speaker.

The third class is the continuous speech recognition system. It allows the user to utter the message in a relatively (or completely) unconstrained manner. Deller, Proakis and Hansen (1993), pointed out that such a recognizer must be capable of performing well in the presence of all the co-articulatory effects. Developing these continuous speech recognition systems is, therefore, the most difficult task. Lee(1989) attributes this difficulty to the following properties of continuous speech:

1. word boundaries are unclear in continuous speech; and
2. co-articulatory effects are much stronger in continuous speech.

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 naive users, who are neither able nor willing to speak words in isolation and clearly, interact with the recognizer.

Vocabulary size: 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, Proakis and Hansen 1993).

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 error.

Most speech recognizers need their users to use only the words of their limited vocabulary, which is not as natural as human communication. But the ultimate goal of speech recognition is and should be to recognize naturally spoken speech. This requires the vocabulary of the system to be comparable to the human vocabulary. To be useful in the natural world and in different activities of human life, speech recognition systems should be able to handle very large vocabularies. One important research direction is to develop ASR systems with open vocabulary (Tang 2005).

Linguistic constraints: Most, if not all, of the present speech recognition systems are unable to reliably determine the identity of a speech input (a phone or a word) based on the speech signal alone. To improve reliability, linguistic constraints are put on a recognizer by using a language model and a pronunciation dictionary. They capture syntactical and lexical constraints, respectively. The more constrained the rules of a language in the recognizer, the less freedom of expression the user has in constructing spoken messages. The challenge of language modeling is to balance the minimization of the search space and maximization of users' freedom of expression.

The pronunciation dictionary determines how the smallest units of recognition form words. For example, it contains the sequence of recognition units for every word of the vocabulary. Small vocabulary speech recognition systems generally do not rely heavily on language models and pronunciation dictionaries to accomplish their tasks. A large vocabulary speech recognition system, however, is dependent on linguistic knowledge included in the input speech.

Type of speech: 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, restarts, incomplete sentences, laughter, coughing, unlimited vocabulary and reduced pronunciation quality.

The primary difference in recognition error rates between read and spontaneous speech are due to dis-fluencies in spontaneous speech (Junqua and Haton 1996). Dis-fluencies in spontaneous speech can be characterized by pause-fillers, word fragments, overly stressed function words, overly long pauses and mispronunciations. In spontaneous speech, people often lack what to say and think along the way. This behavior provokes an interruption of the speech flow, followed by a restart of the utterance. Such interruptions are filled with human (e.g. Lip smacks, laughter) and non-human (non-articulatory) noises which need to be modeled in speech recognition (Schultz and Rogina, 1995). In spontaneous speech, the duration of a phone is usually shorter than in read speech.

Spontaneous speech is, therefore, both acoustically and grammatically difficult to recognize. Using spontaneous speech data in the training phase of a continuous speech recognizer helps to model spontaneous speech effects and improves recognition performance. But the preparation of a spontaneous speech corpus is much more difficult and expensive than the preparation of a read-speech corpus.

Environment: Speech recognizers may require the speech to be clean from environmental noises, acoustic distortions, microphone and transmission channel distortions, or they may ideally handle any of these problems.

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 fluorescent lighting in the office; heavy equipment noise in a factory environment; or cockpit noise in an aircraft. The noise might also be created by the speaker himself in a form of lip smacks, breath takes, pops, clicks, coughs, or sneezes.

The essential difference between controlled conditions and actual world environments is the mis-match introduced between training and testing. In the case of noisy speech, experiments have shown that a system trained in conditions, with respect to the Signal-to-Noise Ratio (SNR) and the type of the noise, similar to those used during testing gave good recognition performance.

In summary, it is much easier to develop speech recognition systems with constraints like limiting the vocabulary size; requiring isolated word type; filtering environmental noise; limiting only to speakers who train them; with no cognitive abilities to learn from mistakes than without any constraints. In contrast, most natural application domains that would benefit from speech recognition, neither use discrete, clearly articulated utterances from a single person in a quiet environment, nor is it generally possible to have the system trained by its user population.

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).

3.4 Approaches to the Development of ASR

Rabiner and Juang (1993) categorized approaches in speech recognition into three broad categories: Acoustic-phonetic, Artificial intelligence and Pattern recognition.

As the name implies, the first approach bases itself on the theory of Acoustic-phonetics. Acoustic-phonetics also studies the physical qualities of the sound wave form which is emitted by the speaker and perceived by the hearer. To be specific, Acoustic phonetics examines the frequency, amplitude and duration of the sound wave passing between speaker(s) and hearer(s).

The Acoustic-phonetic approach assumes that the phonetic units are broadly characterized by a set of features, such as formant frequency, voiced/unvoiced, and pitch. These features are extracted from the speech signal and are used to segment and label the speech.

The process of recognition in this approach involves three steps (Rabiner 1999):

> *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 step is a segmentation and labeling phase in which the speech signal is segmented into stable acoustic regions, followed by attaching one or more phonetic labels to each segmented region, resulting in a phone lattice characterization of the speech.*

> *The last step attempts to determine a valid word (or string of words) from the phonetic label sequences produced by the segmentation and labeling.*

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

1. The method requires extensive knowledge of the acoustic properties of phonetic units.
2. The choice of features is mostly based on ad hoc considerations.
3. The design of sound classifiers is not optimal because the choice of features in the use of Classification And Regression Tree (CART) methods is most likely to be sub-optimal.
4. No well-defined automatic procedure exists for tuning the method on real labeled speech.

The other approach in speech recognition that is competing with the well performing pattern recognition approach is the Artificial Intelligence (AI) approach. Rabiner and Juang (1993) take the AI approach as a hybrid of the Acoustic-phonetic and Pattern-recognition approaches. It is an attempt to automate human intelligence in visualizing, analyzing and decision making based on measures of acoustic features. Rabiner and Juang (1993), also pointed out that the basic idea of the Artificial intelligence approach to speech recognition 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 AI approach incorporates knowledge about the world and the background of the speech into the ASR system.

According to Rabiner and Juang (1993), 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 (phonemic, lexical, etc.) 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.

Artifical neural networks (ANN) can be considered as an implementational architecture for the other basic approaches, (Rabiner and Juang 1993), especially the pattern recognition approach which usually is implemented using HMMs. Neural networks have been applied in the area of speech recognition much later than HMM. Today, however, they are used to develop hybrid HMM/ANN speech recognizers.

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. The pattern recognition approach requires no explicit knowledge of speech. This approach has two steps, namely, training of speech patterns based on some generic spectral parameter set and recognition of patterns via pattern comparison.

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.

The reasons that have made the pattern recognition approach so popular are:

- the underlying statistical (mathematically precise) framework;
- the ease and availability of training algorithms for estimating the parameters of the models from finite training sets of speech data;

- the flexibility of the resulting recognition system where one can easily change the size, type, or architecture of the models to suit particular words, sounds etc.; and
- the ease of implementation of the overall recognition system.

The most successful and popular method of the pattern recognition approach in the area of speech recognition is the Hidden Markov Model (HMM). 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 $N \times N$ stochastic matrix defining transitions between states whose elements $a_{ij} = P(\text{state } j \text{ at time } t \mid \text{state } i \text{ at time } t - 1)$, are the transition probabilities.

Its output symbols are probabilistic, and all symbols are possible at each state, or transition with their own probability. Therefore, each state or transition is associated with a probability distribution of all possible symbols. (Some researchers associate the output probability with the states, while others associate it with the transition. But this has no significant effect on the behavior of the model.) In other words, 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 mathematical framework of the HMM method enables us to combine modeling of stationary stochastic processes (for the short time spectra), and the temporal relationship among the processes, (via a Markov chain) together in a well defined probability space.

In addition, this combination of short time static characterization of the spectrum within a state, and the dynamics of change across states, allows the decomposition of the measure of the observation probability, given the model into a summation of the joint probability of the observation and the state sequence. The decomposition permits an independent study and analysis of the behavior of the short-time processes and the long-term characteristic transitions.

Another advantage of HMM comes from the fact that it is relatively easy and straightforward to train a model from a given set of labeled training data.

Flexibility is also an attractive feature of the basic HMMs. 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 have also different alternatives of HMM topologies with different numbers of states. It is also possible to build HMMs that can de-code speech in various hierarchies that range from phones to sentences. Some of these features of HMMs are presented later in this section.

These strengths have made HMMs the predominant method in current automatic speech recognition technology and research.

An HMM has the following basic mathematical elements:

1. N - the number of states in the model. The individual states can be denoted as $\{1, 2, \cdots, N\}$, and the state at time t as q_t
2. M - the number of distinct observation symbols per state. If the observations are continuous then, M is infinite.
3. The state-transition probability distribution $A = \{a_{ij}\}$ where

$$a_{ij} = P[q_{t+1} = j | q_t = i], \qquad 1 \leq i, j \leq N \tag{3.3}$$

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

$$b_j(k) = P[o_t = v_k | q_t = j], 1 \leq k \leq M, \tag{3.4}$$

defines the symbol distribution in state $j, j = 1, 2, \cdots, N$. If the observations are continuous, then we will have to use a continuous probability density function, instead of a set of discrete probabilities. In this case, we specify the parameters of the probability density function. Usually the probability density is approximated by a weighted sum of M Gaussian distributions $\mathcal{N}$.

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm}\mathcal{N}(\mu_{jm}, \Sigma_{jm}, o_t), \qquad c_{jm} \leq 0, 1 \leq j \geq N, 1 \leq m \geq M, \tag{3.5}$$

where,

c_{jm} = weighting coefficients

μ_{jm} = mean vectors

Σ_{jm} = Covariance matrix

c_{jm} should satisfy the stochastic constraints, $c_{jm} \geq 0, 1 \leq j \geq N, 1 \leq m \geq M$ and $\sum\limits_{m=1}^{M} c_{jm} = 1, \qquad 1 \leq j \geq N$

5. The initial state distribution $\pi = \{\pi_i\}$ in which

$$\pi = P[q_1 = i], \qquad 1 \leq i \leq N \tag{3.6}$$

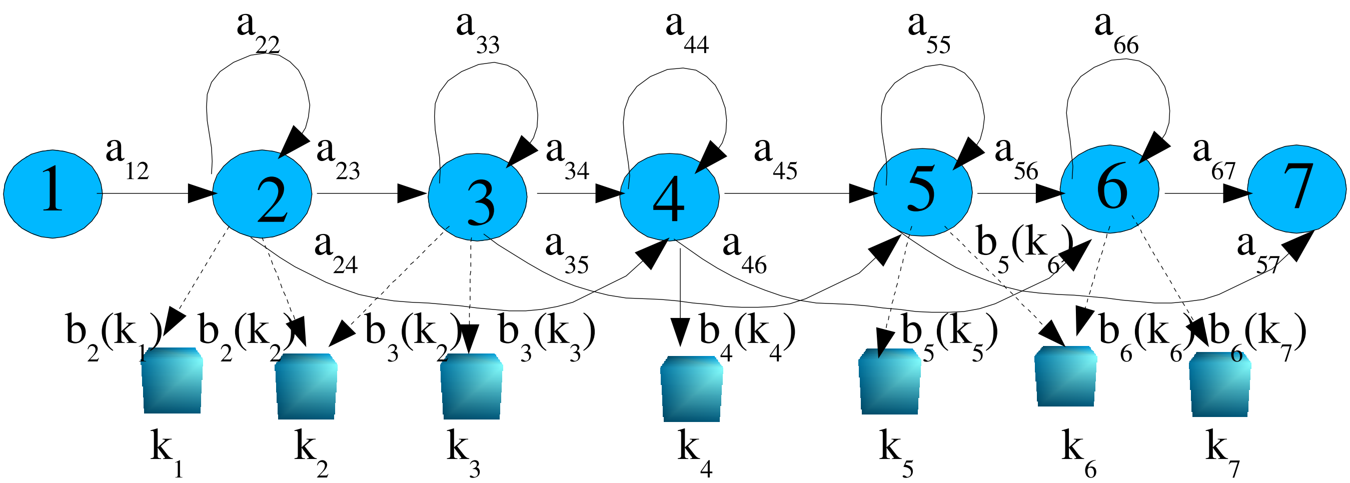

Figure 3.2: An Example of HMM

The HMM model in Figure 3.2 is a model of five emitting states that are numbered from 2 to 6, and have 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 a 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 first-order hidden Markov models, they reduce the number of parameters, and also make learning and decoding algorithms extremely efficient (Lee 89).

3.5 Large Vocabulary Continuous Speech Recognition

Rabiner and Juang (1993) pointed out that the standard approach to large vocabulary continuous speech recognition is to assume a simple probabilistic model of speech production, in which a specified word sequence W produces an acoustic observation sequence Y with probability $P(W_i|Y)$, where W_i is the i^{th} vocabulary word. Using Bayes' rule, the maximum of $P(W_i|Y)$ can be computed by:

$$\hat{W} = arg\max_i P(Y|W)P(W) \tag{3.7}$$

$P(Y|W)$ in this equation is computed by the acoustic model, and P(W) is computed by the language model.

3.5.1 Designing a Hidden Markov Model

Designing an HMM for speech modeling means determining the type of the model. An HMM can be classified on the basis of the structure in its transition matrix, type of its observation symbol and the number of states.

As indicated by Rabiner and Juang (1993), on the basis of the structure in the transition matrix, HMM can be classified as ergodic models and left-to-right (Bakis) models. Ergodic model is an HMM in which "every state of the model can be reached from every other state in a finite but aperiodic number of steps". This kind of model has a property that all transition co-efficients are positive.

Rabiner and Juang (1993) characterize left-to-right model as follows:

- As time increases, the state index increases, or (stays the same). This can be described mathematically as:

$$a_{ij} = 0, \qquad\qquad j < i \tag{3.8}$$

- It has a well-defined initial and final state, which is:

$$\pi_i = \begin{cases} 0, & i \neq 1, \\ 1, & i = 1 \end{cases} \tag{3.9}$$

  that is forcing the state sequence to begin in state 1 and end in the last state.

- Large changes in the state indices do not occur. This is achieved by limiting Δ_i in the following equation to 2. This is the characterizing feature of Bakis model.

$$a_{ij} = 0, \qquad\qquad j > i + \Delta_i \tag{3.10}$$

The model topology that is generally adopted for speech recognition is a left-to-right or Bakis model because the speech signal varies in time from left to right (Deller, Proakis and Hansen 1993).

Within each state of the HMM, the spectral vector is represented by either a discrete density (i.e. a distribution over a spectral code-book) or a continuous density (e.g. mixture Gaussian), or a mixed density (i.e. a continuous density over a code-book of common spectral shapes) (Lee, Rabiner and Pieraccini 1992). Discrete density HMM is an HMM in which the observations are considered to be discrete and consequently, the discrete probability density is used. Since speech observations are continuous in nature, various methods have been used to discretize the speech signal. One of these methods is vector quantization. However, there might be a serious degradation of speech signal associated with such a quantization. It is, therefore, advantageous to be able to use continuous density HMM (Rabiner and Juang 1992).

The continuous density method describes the spectral density of each state for every sub-word unit in terms of a mixture of Gaussian densities. Each mixture component has a spectral mean and variance which is highly dependent upon the spectral characteristics of the sub-word unit. To train this kind of HMM for every sub-word unit may require a very large training corpus. When the available corpus is not large enough, different problems arise. They can, however, be alleviated by sharing distributions among transitions of different models.

3.5.2 Acoustic Models

As we have pointed out earlier in this chapter, the most popular and successful approach to develop an acoustic model is the Hidden Markov Model (HMM). Developing an HMM usually involves the following steps (Young et. al 2002):

1. define a set of L sound classes for modeling, such as phones or words; call the sound classes;
$$V = \{v_1, v_2, \cdots, v_L\}; \tag{3.11}$$

2. for each class, collect a sizeable set (a training set) of labeled utterances that are known to be in the class;

3. based on each training set, solve the estimation problem to obtain a "best" model λ_i for each class $v_i, i = 1, 2, \cdots, L$;

4. during recognition, evaluate;

$$P(O|\lambda_i), \quad i = 1, 2, \cdots, L, \tag{3.12}$$

for the unknown utterance O and identify the class v_i that produced O if

$$P(O|\lambda_i) = \max_{1 \leq i \leq L} P(O|\lambda_i) \tag{3.13}$$

In the development of the Hidden Markov Model methodology, the following problems are of particular interest:

1. Given the observation sequence $O = (o_1 o_2 \cdots o_T)$ and a model $\lambda = (A, B, \pi)$, how do we efficiently compute $P(O|\lambda)$?
2. Given the observation sequence $O = (o_1 o_2 \cdots o_T)$, and the model λ, how do we choose a corresponding state sequence $q = (q_1 q_2 \cdots q_T)$ that is best to "explain" O?
3. How do we adjust the model parameters $\lambda = (A, B, \pi)$, to maximize $P(O|\lambda)$?

Evaluation is the first problem, and can be viewed as a matter of scoring how well a particular model matches a given observation sequence. The solution to this problem lies in the *Forward procedure*, which computes the probability of the partial observation sequence, $o_1 o_2 \cdots o_t$, until (time t) and state i at time t, given the model λ. It can be solved inductively as follows:

1. Initialization

$$\alpha_1(i) = \pi_i b_i(o_1) \qquad\qquad 1 \leq i \leq N \tag{3.14}$$

2. Induction

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{i,j}]\, b_j(o_{t+1}), \quad 1 \leq t \leq T-1, \quad 1 \leq j \leq N \tag{3.15}$$

3. Termination

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{3.16}$$

The second problem is related to uncovering the "hidden" part of the HMMs, i.e., finding the optimal state sequence. The best solution to this problem is the use of the *Viterbi Algorithm* in order to find the optimal state sequence. The whole procedure of the algorithm can be stated as follows:

1. Initialization:

$$\delta_1(i) = \pi_i b_i(o_1), \quad 1 \leq i \leq N \tag{3.17}$$

$$\psi_1(i) = 0 \tag{3.18}$$

2. Recursion:

$$\delta_1(j) = \max_{1 \leq i \leq N}[\delta_{t-1}(i)a_{ij}]b_j(o_t), \quad 2 \leq t \leq T, \quad 1 \leq j \leq N \tag{3.19}$$

$$\psi_t(j) = \underset{1 \leq i \leq N}{\mathrm{argmax}}[\delta_{t-1}(i)a_{ij}], \quad 2 \leq t \leq T, \quad 1 \leq j \leq N \tag{3.20}$$

3. Termination:

$$P^* = \max_{1 \leq i \leq N}[\delta_T(i)] \tag{3.21}$$

$$q_T^* = \underset{1 \leq i \leq N}{\mathrm{argmax}}[\delta_T(i)] \tag{3.22}$$

4. Path back-tracking:

$$q_t^* = \delta_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, \cdots, 1 \tag{3.23}$$

The third problem is optimizing the model parameters to best describe how a given observation sequence comes about, as well as developing the recognizer through training the HMMs. The *forward-backward algorithm*, also called the *Baum-Welch algorithm*, does this optimization efficiently. It is the combination of the forward and backward procedures.

We have seen that the forward procedure computes the probability of the partial observation sequence, until (time t) and (state i at time t), given the model λ. The backward procedure gives the probability of the remaining partial observation sequence from $t+1$ to the end, given (state i at time t) and the model λ. It can be solved recursively as follows:
Initialization:

$$\beta_T(i) = 1, \qquad 1 \leq i \leq N \tag{3.24}$$

Induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad t = T-1, T-2, \cdots, 1; \quad 1 \leq i \leq N \tag{3.25}$$

To describe the procedure of the forward-backward re-estimation, let us define two other variables that are required:

1. The probability of being in state S_i at time t and in state S_j at time $t+1$ as:

$$\xi_t(ij) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda) \tag{3.26}$$

   It can be defined in terms of the α and β as:

$$\xi_t(ij) = \frac{\alpha_t(j) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{i=1}^{N} \alpha_t(j) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)} \tag{3.27}$$

2. The probability of being in state S_i at time t given the observation sequence and the model as:

$$\gamma_t(i) = \frac{\alpha_(i) \beta_t(i)}{\sum_{i=1}^{N} \alpha_(i) \beta_t(i)} \tag{3.28}$$

   These two variables, ξ and γ, can be related with one another as follows:

$$\gamma_t(i) = \sum_{i=1}^{N} \xi_t(i, j) \tag{3.29}$$

It is now possible to re-estimate parameters of $\bar{\lambda}$ $(\bar{\pi}, \bar{A}, \bar{B})$ using ξ and γ that are computed from the initial values of the model parameters as follows:

$$\bar{\pi} = \gamma_1(i) \tag{3.30}$$

$$\bar{A} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{3.31}$$

and,

$$\bar{B} = \frac{\sum\limits_{\substack{t=1\\ O_t=v_k}}^{T} \gamma_t(j)}{\sum\limits_{t=1}^{T} \gamma_t(j)} \tag{3.32}$$

Eventually new parameters of λ will be used to compute ξ and γ which will be used to compute another parameters of λ until $\bar{\lambda}$ cannot be improved further, which is the point of convergence.

We note that training HMM models starts with initialization. The initialization of the model for a set of sub-word HMMs prior to re-estimation can be achieved in two different ways. A small set of hand-labeled bootstrap training data can be used to initialize each sub-word HMM individually. A second and simpler initialization procedure is to assign the global speech mean and variance to every Gaussian distribution in each sub-word HMM. This so-called flat start procedure implies that during the first cycle of embedded re-estimation, each training utterance will be uniformly segmented. The hope is that enough of the sub-word models align with actual realizations of that sub-word so that in the second and subsequent iterations, the models align as intended.

In training sub-word HMMs, the parameters of continuously spoken utterances are used as an input source of training data to re-estimate the complete set of sub-word HMMs simultaneously. A transcription, in terms of sub-word units, is needed for each input utterance. Then all of the sub-word HMMs corresponding to the sub-word list are joined together to make a single composite HMM. This composite HMM is used to collect the necessary statistics for the re-estimation. When all of the training utterances have been processed, the total set of accumulated statistics is used to re-estimate the parameters of all of the sub-word HMMs. It is important to emphasize that in this process the transcriptions are only needed to identify the sequence of sub-words in each utterance. No boundary information is required.

The major problem with HMM training is that it requires a great amount of speech data. To overcome the problem of training with insufficient speech data, a variety of sharing mechanisms can be implemented. For example, HMM parameters are tied together so that the training data is pooled and more robust estimates result. The parameter tying is often involved in the incremental manipulations of the HMM set. We can also restrict the model to a variance vector for the description of output probabilities, instead of a full covariance matrix. Rabiner and Juang(1993) pointed out that for the continuous HMM models, it has been found that it is preferable to use diagonal covariance matrices with several mixtures, rather than fewer mixtures with full covariance matrices to perform reliable re-estimation of the components of the model from limited training data.

3.5.3 Language Model

A language model describes how the text in a given corpus is organized at the word-level. It provides the probabilities that word W (e.g. ice) is followed by word W1 (e.g. cream) in any given text. A language model generally uses statistical concepts of probability to describe the likelihood of W1 following W, but other models, such as finite-state models, can be used independently or together with statistical models.

The goal of the statistical language model is to provide an estimate of the probability of a word sequence W for the given recognition task. The ideal situation is that a language model estimates the probability of all the possible sequences or a sequence of any number (N) of words in a language. In practice it is only possible to compute the sequence of a maximum of three (trigram language model) or four words, because the (n-gram) models require V^N parameters, where (V) is the vocabulary size. This number very quickly becomes too large. Jelinek(1990) also pointed out that for a vocabulary of size V, there are V^{i-1} different histories, therefore, to specify $P(w_i|w_1, \cdots, w_{i-1})$ completely, V^i values would have to be estimated, which can neither be stored nor retrieved when needed by the recognizer. To overcome this problem, Bahl, et al. (1983), proposed another approach using decision trees to selectively ask questions about the previous 20 words to determine the probabilities of the next word. This approach requires a very time-consuming tree-building process.

A statistical language model (the probability of a word sequence - $P(W)$) has to be trained with a given (large) text corpus. This word sequence probability of (N) words (n-gram) is approximated by:

$$P_N(W) = \prod_{i=1}^{Q} P(w_i|w_{i-1}, w_{i-2}, \cdots, w_{i-N+1}) \tag{3.33}$$

The conditional probabilities,

$$P(w_i|w_{i-1}, \cdots, w_{i-N+1}) \tag{3.34}$$

can be estimated by the simple relative frequency approach,

$$\hat{P}(w_i|w_{i-1}, \cdots, w_{i-N+1}) = F(w_i, w_{i-1}, \cdots, w_{i-N+1})F(w_{i-1}, \cdots, w_{i-N+1}) \tag{3.35}$$

where F is the number of occurrences in the sequence of its argument in the given training corpus. In most speech recognition systems an estimate of only two successive words (bigram language model) is used. This is due to the fact that if a small corpus is used to estimate word sequence probabilities for larger N, most of the numerators become zero or very small. This phenomenon causes the probability estimate of the sequence to be zero or very small. Jelinek(1990) suggested smoothing (ensuring some probability estimate greater than zero or a minimum threshold) as a solution to this problem of zero probabilities. Various smoothing techniques, such as backing-off, linear interpolation, linear discounting, deleted interpolation, maximum entropy, co-occurrence smoothing, and count re-estimation (Roukos 96), have been devised. With the development of computing technology in terms of processing speed and storage space, the use of different techniques of smoothing and larger training corpora, the length of word sequences being available (Ns) in the (n-gram) language model can be increased.

The potential contribution of a language model (as compared to another language model) in the context of speech recognition is measured in terms of its perplexity. Perplexity can be considered to be a measure of, on average, how many equally different and most probable words can follow any given word. Test set perplexity is an important parameter in specifying the degree of sophistication of a recognition task (Rabiner and Juang 1993).

Young et al., (2002) puts its mathematical formula, which they derive from the equation of entropy, as follows:

$$PP = \hat{P}(w_1, w_2, \ldots, w_m)^{-\frac{1}{m}} \tag{3.36}$$

where, $\hat{P}(w_1, w_2, \ldots, w_m)$ is the probability estimate assigned to the word sequence (w_1, w_2, $\ldots, w_m$) by a language model.

Perplexity increases with the vocabulary size of the test set. On the other hand, language models of small vocabulary suffer from the problem of high rate of out-of-vocabulary (OOV) words. Byrne et al., (2000) introduced new language models based on morphemes. Using these models, they managed to reduce significantly the vocabulary size of the language model and the rate of OOV, consequently improving the word recognition accuracy.

3.6 Fundamental Sub-word Units

Large Vocabulary Automatic Speech Recognition Systems (LVASRSs) require modeling of speech in smaller units rather than words because the acoustic samples of most words will never be seen during training, and therefore, can not be trained. In LVASRSs, the number of whole words is in the thousands and most of them occur very rarely, consequently training of models, for words, is generally impractical. That is why LVASRSs require a segmentation of each word in the vocabulary into sub-units that occur more frequently and can be trained more robustly than words. It is also possible to deal with words unseen during training since they can just be decomposed into the subunits. As a word can be decomposed into sub-units in different ways, there is a need to choose the most suitable sub-unit that fits the purpose of the system.

There are two alternatives for choosing the fundamental sub-word units, namely acoustically-based and linguistically-based units (Lee, Rabiner and Pieraccini 1992). They pointed out that acoustic units are the labels assigned to acoustic segment models, which are defined on the basis of procuring a set of segment models that spans the acoustic space determined by the given, unlabeled training data. The linguistically-based units include the linguistic units, e.g. phones, demi-syllables, syllables and morphemes.

It should be clear that there is no ideal (perfect) set of sub-word units. Although the phones are very small in number and relatively easy to train, they are much more sensitive to contextual influences than the larger units. Other extremes are the syllables which are the longest units and the least context sensitive ones. However, these are too many in a number of languages, such as English, to be trained properly. Consequently we need to look for a sub-word unit that is limited in number and is as large in size, as possible. Thus researchers in English ASR are led to choose phones and leads us to consider syllables as an alternative for the development of Amharic ASR, because Amharic has only 233 distinct CV syllables. As presented in Table 3.1, we can compare these sub-word units for English and Amharic, with respect to their coverage of language, context sensitivity, ease of initialization, and trainability.

Issues	Phones	English Syllables (20,000)	Amharic Syllables(233)
Coverage of language	Good coverage	Less coverage	Good coverage
Context sensitivity	Context sensitive	Less context sensitive	Less context sensitive
Ease of initialization	Very easy	Difficult	Easy
Trainability	Trainable	Less trainable	Trainable

Table 3.1: Issues of Sub-word Selection

A syllable usually consists of a vowel surrounded by one or more consonants. The syllable is a sub-word unit with a longer duration that enables us to simultaneously exploit temporal and spectral variations for the development of sub-word HMM. Moreover, the syllable has a close connection to articulation, integrates some co-articulation phenomena, and has the potential for a relatively compact representation of conversational speech. The problem with syllables in many of languages is their large number which is not the case in Amharic.

The most prominent approach to the problem of developing LVCSR is the use of phones for modeling spoken words. However, it has been known that phones are too small to model temporal patterns and variations in continuous speech. The use of triphones has, therefore, become the dominant solution to the problem of the context sensitivity of phones.

Triphones model the dependence on both the right and left context of a phone. They are models of a single phone conditioned on its immediate neighbors. Triphone models can be constructed either word-internally, or across-words. When constructing word-internal models, context beyond the word borders is not considered. On the other hand, for cross-word triphones, the phones at the end or beginning of neighboring words are considered to affect the phone at the beginning or the end of a word, respectively.

Triphones also are relatively inefficient sub-word units due to their large number. Moreover, since a triphone unit spans a short time-interval, it is not suitable for the integration of spectral and temporal dependencies.

That is why a need exists for another unit, e.g. syllable, that is capable of exploiting both the spectral and temporal characteristics of continuous speech (Ganapathiraju et. al. 1997).

Ganapathiraju et. al. (1997) have explored techniques to accentuate the strengths of syllable-based modeling with a primary interest of integrating finite-duration modeling and monosyllabic word modeling. Wu et. al. (1998) tried to extract the features of speech over the syllabic duration (250ms), considering syllable-length interval to be 100-250ms. Hu et. al. (1996) used a pronunciation dictionary of syllable-like units that are created from sequences of phones for which the boundary is difficult to detect. Kanokphara (2003) used syllable-structure-based triphones as speech recognition units for Thai.

3.7 The Hidden Markov Model Toolkit (HTK)

Almost all of the information for the description of the toolkit is extracted from the HTK Book (Young et. al., 2002).

The Hidden Markov Model Toolkit (HTK) is a portable toolkit for building and manipulating hidden Markov models using both continuous density Gaussian mixtures and discrete distributions, especially for speech recognition research. Thus, much of its support is dedicated to this task. 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.

HTK enables both methods of model initialization that are mentioned earlier in this chapter. 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 reestimation 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 utterances 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.

There is also an HTK tool that compares the recognizer output with the correct reference transcriptions and counts substitution (S), deletion (D) and insertion (I) errors. The tool gives sentence level and word level accuracy. The former is computed based on the total number of label files which are identical to the transcription files using equation 3.37, while the latter is computed based on the matches between the label files and the transcriptions using equation 3.38.

$$Correct = \frac{H}{N} \times 100 \tag{3.37}$$

where (N) is the total number of label files and (H) is the total number of correctly recognized sentences.

$$Accuracy = \frac{H - I}{N} \times 100 \tag{3.38}$$

where (H) is the total number of correctly recognized words and (N) is the total number of labels in the reference transcription files.

Chapter 4

Corpus Preparation

4.1 Introduction

A speech corpus is one of the fundamental requirements for any speech recognition research. That is the main reason for the preparation of an Amharic speech corpus as a first step in exploring the possibilities of developing an Amharic speech recognition system.

Speech corpus in the scope of this work is defined as 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 the data in-house, or by scientists in other organizations (Gibbon et. al 1997). This chapter gives an overview of speech corpus preparation, a description of the Amharic speech corpus developed for this work and the process of its preparation.

Schiel et. al (2003) mention the following styles of speech that differentiate a speech corpus from others:

- Read 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 project on its own and handled on the basis of an agreement between corpus producer and corpus users. However, in cases like this study where the required corpus is not available, speech recognition experiments are conducted on the newly produced corpus. The advantage in the latter case is that the corpus is produced with full and specific knowledge of its intended use.

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 the first reason, a read speech corpus is prepared and used for this work.

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.

4.2 Preparation of Amharic Speech Corpus

The Amharic speech corpus 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 vocabulary). The Amharic corpus has, therefore, been made to contain the same components. The following is a brief description of the corpus preparation:

The training set is intended to collect speech data for the training of a recognizer. The speaker adaptation set contains speech data, for the purpose of speaker adaptation using pre-recorded speech data. The development test sets are used to test the process of developing the recognizer and identifying the problems, so that the performance of the recognizer is improved through correction of the identified problems. For the purpose of a final evaluation of the recognizer, the evaluation test sets are provided.

4.2.1 The Procedure of Text Selection

The selection of sentences from a text database aims at both a phonetically rich and balanced collection of sentences with regard to the relative frequencies of the sub-word units to be modeled (phones, triphones and syllables). To accomplish phonetic richness, we have selected sentences which contribute to the inclusion of all Amharic syllables. Phonetic balance of the corpus is achieved by selecting those sentences which contribute to the preservation of the distribution of syllables in the language.

Radova and Vopalka (1999) suggested a method of phonetically balanced sentence selection, for read speech corpus design with regard to triphones. Their method is used in our work, with a different sub-word unit - syllables instead of triphones - with more enforcements enriching the frequencies of rare syllables to have them appear at least twenty times. There is, therefore, a need to have a text database that shows the natural frequency of Amharic syllables, and from which usable sentences can be selected.

However, in contrast to other languages like English, there are no easily available electronic text sources for Amharic usable for this work. Since keying in is very expensive, accessible electronic sources of Amharic text have been identified. The archive of the Internet source at the previous Ethiopian News Headlines (ENH), the current EthioZena was deemed usable. We requested for its access and was made available to us online in an encoding called SERA. To interface it with other processing components of this work the text was converted to ethiop encoding (Beyene 1997). Over 100,000 sentences have been acquired from this archive. During the use of the obtained text, some problems were encountered and solved. To mention a few:

- Spelling and grammar errors have been corrected[1];
- Abbreviations have been expanded and foreign (especially English) words have been removed;
- Missing labio-velar Amharic CV syllables have been included;
- Numbers have been textually transcribed;
- Lots of concatenated words have been separated;
- The text has been converted from HTML to plain text, from SERA to ethiop as well as to the convention used in the current project.

The purpose of having a speech corpus for a language is to have a data source used to train a speech recognizer covering as many sub-word units of the language as possible. Ideally, all of them. It should therefore contain speech data required to train a model for each of these units. We tried to achieve this goal by selecting a set of phonetically rich sentences that included all Amharic CV syllables. The natural distribution of these recognition units in the language is also preserved in the corpus by means of selecting phonetically balanced sentences.

To avoid elongated sentences that create difficulty for the readers, sentences with a maximum of twenty words in length have been chosen from the available electronic text

[1] Please note that there is no Amharic spell- and/or grammar checker

resources. At the same time, all of the syllables that are in the archive are included in the selected text. The selected text contains a total of 72,000 sentences. These selected sentences have been manually checked for grammaticality, spelling, foreign words, abbreviations, etc. and have been manually corrected.

From this text database, fifty three sentences have been selected to create a phonetically rich and balanced text database for the speaker adaptation set. From the remaining data in the text database, 10,000 sentences have been selected to create a phonetically rich and balanced database for the training set.

The selection of the sentences that best contribute to build a phonetically rich and balanced text database was automatically done in two steps:

First, the so-called important sentences (Radova and Vopalka 1999) have been selected. These sentences were intended to include all of the syllables that are in the text database. This was completed in two steps:

- step 1, a sentence that contains the largest number of distinct syllables has been selected; and
- step 2, a sentence is selected if it adds the largest number of new syllables to the list of syllables generated in step 1;
    - step 2 is repeated until all of the syllables in the text database are included in the selection.

Second, sentences that enrich the phonetic balance, (their natural frequency in the database) of the recording sets are selected based on an add-on procedure. Based on the distribution of syllables in the current selection, a score, which estimates the utility of a sentence for achieving the desired target distribution, is computed (Radova and Vopalka 1999). The sentence with the highest utility is chosen next, and deleted from the database.

The problem with a medium sized read speech corpus which has to be phonetically rich and balanced is that rare recognition units will be too few in order to train their HMMs properly. Moreover, due to pronunciation errors made by the readers during recordings, the frequency of rarely used syllables decreases even more. As a result, in the selected sentences, nineteen syllables were missing completely, and a few others have too low frequency to be used in a statistical method such as the Hidden Markov Modeling. We solved this problem by collecting Amharic words that contain missing and rare syllables and are active in modern Amharic. We did this in consultation with language experts. Sentences have been constructed using these words according to the grammar of the language. The frequency distribution of the syllables in the set of the selected sentences and the modified speech corpus, is given in Figure 4.1, page 62 and Table 4.1[2], page 63.

The evaluation and development test sets were selected from the remaining sentences in the database. The restriction for the selection of these sets was the vocabulary size. Evaluation and development test sets of 5,000 words contain a minimum of 5,000 different words, while evaluation and development test sets of 20,000 words contain a minimum of 20,000 different words. That is, sentences have been randomly selected until the intended vocabulary size was achieved. For example, when sentences were selected for the evaluation test set of 5,000 words, a sentence was first selected randomly and the words in the sentence are counted. Another sentence was added, and the word statistics was updated. This was repeated until the word count was greater than 5,000.

4.2.2 The Speech Recording

Having the text corpus, the next important step in read-speech corpus preparation is recording the speech. In the recording of the selected sentences, the speaker is asked to read exactly what is presented to him or her. The text to be read may either be printed on paper or presented on a computer screen.

In read speech recordings the degree of control is very high. For example, during the

[2] This Table shows only syllables of high and low frequency

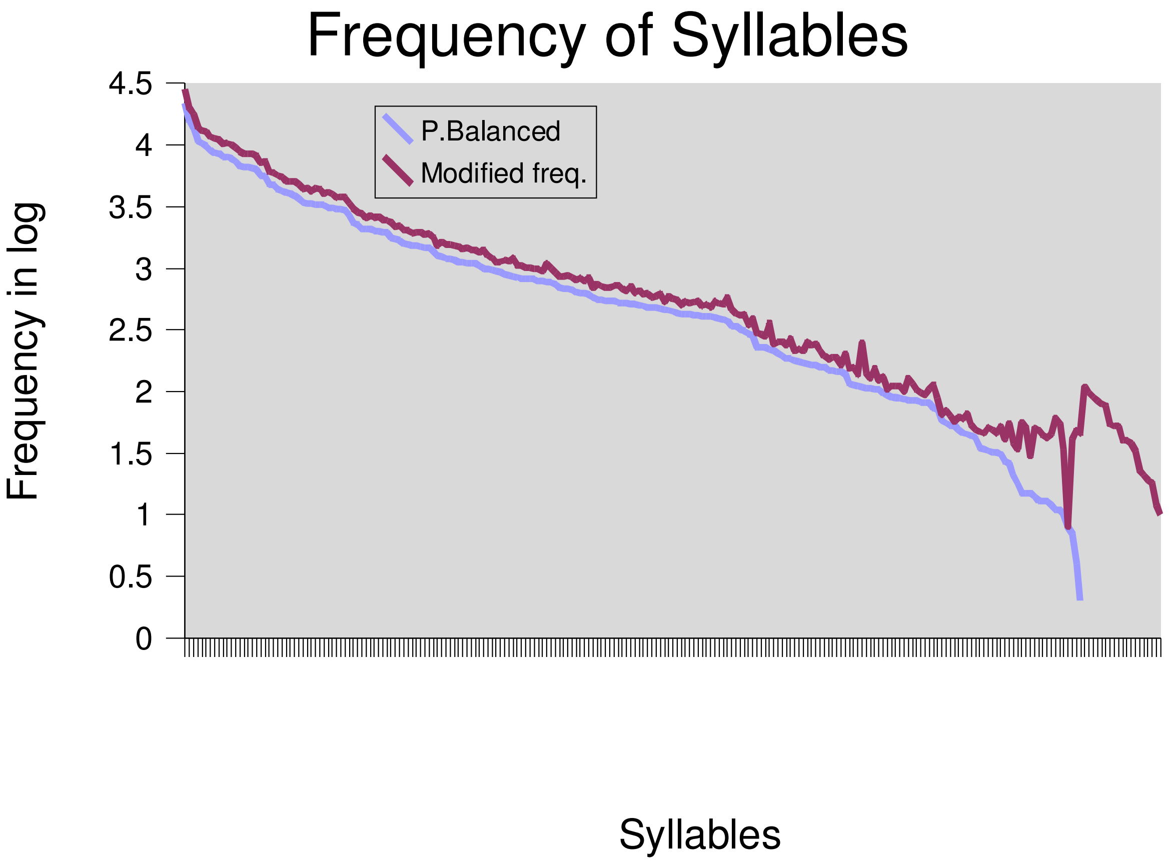

Figure 4.1: Frequency of Syllables

recording, each utterance of the speaker can be checked directly for errors, and, if an error is found, the speaker is asked to re-read the text (Gibbon et. al 1997).

The recording of the Amharic speech corpus has been done in Ethiopia, in an office environment. The text has been read by 124 native Amharic speakers. For the recording, a headset close speaking microphone was used. A brief description of the recording procedure is as follows:

For the recording purpose, a program has been written in Perl that displays one sentence for the speaker to read. The whole recording was done in the presence of the researcher. The speaker was first explained the purpose of the project and instructed what to do. The recording session was controlled by the researcher. The control included: running the recording program, starting the recording session, breaking the recording, playing back

Syllable	Phonetically Balanced Frequency	Modified Frequency	Syllable	Phonetically Balanced Frequency	Modified Frequency	Syllable	Phonetically Balanced Frequency	Modified Frequency
ne	21954	28254	na	5564	7387	yi		79
te	15509	19877	mA	4788	6040	kue		77
we	13797	17877	bA	4771	5938	hue		54
se	10846	13957	ye	4352	5651	que		52
yA	10339	13016	mi	4247	5463	gui		52
ya	9790	12785	ca	4092	5087	guE		40
ta	8996	11642	ge	4029	5044	Pa		40
ba	8584	11309	rA	3873	5058	qui		38
Ha	8513	11138	lA	3652	4795	kui		33
He	8032	10178	...	...	...	Pi		23
le	8005	10317	vu	10	34	hua		21
la	7773	10095	vuA	8	8	quE		19
re	7301	9527	ZuA	7	41	kuE		18
ma	6686	8810	xi	4	48	huE		12
ce	6619	8459	Po	2	46	Pu		10
nA	6605	8413	kua		109			
me	6479	8449	gua		98			
da	6343	8125	gue		90			
ga	5621	7155	qua		84			

Table 4.1: Improvement in Syllable Frequency

the recorded speech, re-recording the sentence (if required), moving to the next sentence. Every speaker was instructed to only start the recording when she or he was ready. The other controls were done by the researcher himself. After the entire session for a reader was finished, all the utterances were listened to both by the reader and the researcher for corrections.

The recorded training set consists of a total of 10850 different sentences. It includes 450 sentences that are necessary to consider missing syllables and enrich the frequency of the rare ones. The training set was read by eighty speakers of the Addis Ababa dialect, seventy of them read 100 sentences each and ten of them read 145 sentences each. We have also recorded speech of twenty speakers of the other four dialects, who read 120 sentences (100 from the phonetically balanced training text database, and 20 sentences from the set of sentences that are constructed to increase the frequency of rare syllables) each, for the training set. Table 4.3, 65 shows the age and sex distribution of all speakers. This distribution is similar to the one founed in WSJCAM0 (Frasen 1994). Due to time constraints for recording in the respective regions and the small size of the speaker sample it was difficult to keep the age balance for the dialect speakers.

Test and speaker adaptation sets were read by twenty other speakers of the Addis Ababa dialect and four speakers of the other four dialects. For the 5,000 vocabulary (development and evaluation) and the 20,000 vocabulary (development and evaluation) test sets, 18 and 20 different sentences have been selected, for each speaker, respectively. Table 4.2 shows the number of sentences that have been automatically selected from the text database and read for the collection of speech data for the corpus. The sentences that are intended to include the missing and rare syllables as well as those read by the other four dialects are not included in the Table. In our experiment that is presented in Chapter 5, we used all the training speech data which contains 10850 sentences and corresponds to 20 hour of speech.

Sets	Number of Selected sentences	Number of Recorded sentences
Training set	10000	8000
Speaker adaptation set	53	53
Development test set for 5000 vocabulary	1000	360
Development test set for 20000 vocabulary	4000	400
Evaluation test set for 5000 vocabulary	850	360
Evaluation test set for 20000 vocabulary	3000	400

Table 4.2: Elements of the Amharic Speech Corpus

The average duration of the recording per speaker was 1 hour and 15 minutes for the training set and 1 hour and 30 minutes for the other sets (adaptation and test sets together).

Age Range	Training set		Test sets	
	Male	Female	Male	Female
Speakers of the Addis Ababa dialect				
18-23	18	18	3	3
24-28	12	12	3	3
29-40	5	5	3	3
Older than 40	5	5	1	1
Total	40	40	10	10
Speakers of the other four dialects				
18-23	10	3	4	
24-28	6	1		
Total	16	4	4	
Grand Total	56	44	14	10

Table 4.3: Age and Sex Distribution of the Readers.

4.3 Segmentation of the Corpus

Segmentation is the discrete (categorical) description of a physical signal (coding). It usually consists of a closed set of symbols and a scheme linking these symbols to points, or segments in time.

Segmentation of the speech signal is often done manually by experts. This has two major drawbacks: i) The process is laborious and tedious, requiring e.g., extensive spectrogram reading and listening. It has been estimated that 20-60 minutes would be required to transcribe one minute of speech manually. ii) Due to the lack of an objective criterion, manual procedures unavoidably will exhibit some inconsistencies. Defining a difference in labeling when phoneme boundaries differ by more than 10 ms, a study (Svendsen 1995) reported that 28% of the phoneme boundaries differed when cross-comparing two different labelers and 24% of the boundary positions differed when comparing two segmentations made by the same person. Transcriptions created by automatic methods on the other hand have attractive properties of being consistent and reproducible. However automatically segmented data is not yet as reliable as that of the experts. Therefore, it needs manual checking and in some cases correction.

Several methods have been proposed to speed up this process or to make it fully or partially automatic. The most successful methods have been borrowed from automatic speech recognition, such as HMM (Brugnara et al. 1993) techniques, because automatic alignment can be viewed as a simplified recognition task. An HMM recognizer has been used to do forced alignment, that is, a known sequence of phoneme models has been used with the Viterbi algorithm to generate a phonetic alignment (Kare 2001). However, it was required that an orthographical transcription exists. This has been obtained from the text that is read when the speech was recorded.

The Amharic speech corpus is segmented semi-automatically at a syllable-level. First, a limited set of manually segmented speech data is used to build initial models. These models are then used to perform segmentation of the remaining training material by Viterbi decoding given the correct word-level transcription. This produces the optimal state sequence from which the segmentation can be computed. Next the resulting label files are manually checked by non-expert listeners.

As mentioned earlier, this corpus is a read speech corpus. There is, therefore, an orthographic transcription that can be used for this kind of automatic segmentation, but there were a number of mispronounced words for which the transcription has to be changed according to the pronunciation. This required listening to every speech file. Out of these data, a set of speech files that includes all Amharic syllables with a minimum frequency of 20 has been manually annotated. Using this segmented speech data, HMM models have been trained by the Baum-Welch method for each of the syllables.

Figure 4.2, page 67, shows a part of a segmented and transcribed speech signal with labels that contain the following Amharic syllables: እ ና ን ተ ም መ ቀ በ ሪ ያ እ ን ዳ [?I na nI tA mI mA qA bA ri ya ?I nI da]. The complete sentence of the speech reads as እናንተም መቀበሪያ እንዳታጡ ተጠንቀቁ። [?InanItAmI mAqAbAriya ?InIdatat‘u tAt‘AnIqAqu].

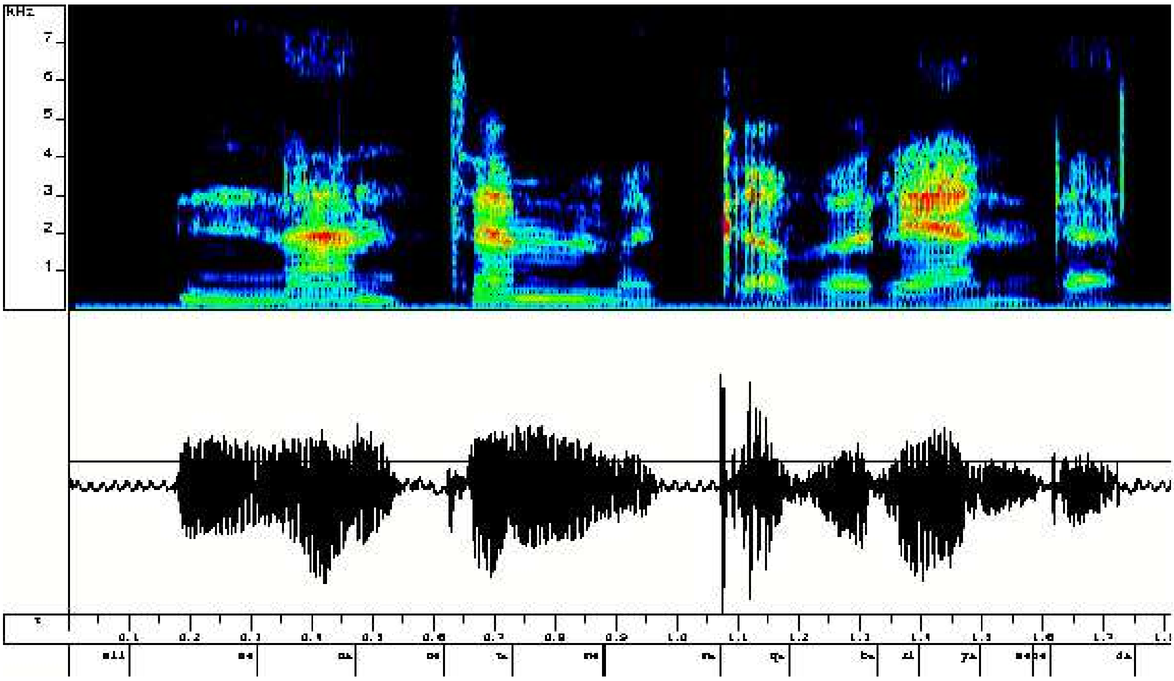

Figure 4.2: An Example of a Segmentation

4.4 Documentation

It is always of crucial importance that the collecting procedure of a speech corpus is documented as elaborately as possible. It is good practice to record all possible details about, for instance, sex and age of speakers, type of speech material, (isolated words, sentences, discourse, etc.), place of recording (in a laboratory, on location, etc.), type of microphone, and recording medium. A well documented speech corpus may also be used for other research.

Gibbon et. al (1997) pointed out that the characteristics of the included speech data have to be described in terms of:

- type of database (speech acoustic wave forms, acoustic data with phonetic labelings, acoustic data with the corresponding orthographic forms, acoustic data with the corresponding phonetic transcription, acoustic data with the corresponding recognition-units transcription, etc.);

- size of data (how many hours/minutes of speech);
- number of speakers and how they are selected, either arbitrarily, or with respect to some characteristics such as sex, age or age ranges, physical state, psychological, experience, attitude, accent, etc.;
- acquisition channels (single microphone, set of microphones, similar telephone handsets, or as many handsets as possible):
- environment conditions (noisy, quiet, all conditions, etc.); and
- many other constraints derived from the operating condition.

A brief documentation of the corpus is given in Appendix A and a detailed one according to the above checklist is a part of the corpus's distribution.

Chapter 5

Automatic Speech Recognition for Amharic

5.1 Introduction

Although lots of research has been conducted in the area of ASR, only few results are available for Amharic. A few students (Martha 2003, Zegaye 2003, Kinfe 2002, and Solomon 2001) at Addis Ababa University made first experimental attempts in the area of speech recognition for Amharic. However, their research was restricted to the recognition of isolated syllables or small vocabularies. Martha (2003) developed word-based ASR for the application of command and control. Zegaye (2003) trained a phone-based LVCSR system using a part of the corpus described in chapter 4, but was not able to carry out a systematic evaluation. Kinfe (2002) conducted his research using a small corpus that has only 170 vocabulary and collected from only fifteen training speakers. Solomon (2001) developed an ASR that recognizes isolated CV syllables.

This Chapter presents our experiments and findings in the development of ASRS for Amharic. The work includes:

- preparation of pronunciation dictionaries using different pronunciation units,
- extraction of speech features,
- development of statistical language models for different test sets,
- development of different syllable- and triphone-based ASR systems using different HMM topologies, and
- evaluation of the recognizers and comparison of their performance.

5.2 Pronunciation Dictionary

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 pronunciations. 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.

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).

We have prepared two canonical Amharic pronunciation dictionaries, each of which transcribes over 50,000 words. One is the transcription of words in terms of CV syllables and the other is the pronunciation of words in terms of phones. All the test pronunciation dictionaries of phone- and syllable-based recognizers have been extracted from these canonical pronunciation dictionaries.

We have also prepared an alternative syllable pronunciation dictionary that transcribes over 25,000 words. It is the by-product of our work on speech segmentation. As it has been presented in Section 4.3, the Amharic speech corpus is segmented automatically at syllable level and edited manually. The manual editing is conducted by repeatedly listening to the speech and correcting the labels that do not match the actual pronunciation. As a result, mispronunciations and other variations are labeled as they have been pronounced. Although the work is time consuming and requires much human resource, its outcome (alternative pronunciation dictionary and correct transcription of the speech with correct time alignment) is very helpful for the development of ASRS. Due to time and financial limitations, the segmentation is conducted only for 8000 sentences of training speech data that contain about 25,000 different words. That is why our alternative pronunciation dictionary has only 25,000 different words.

The inflectional feature of Amharic, as explained in Chapter 2, considerably increased the size of our pronunciation dictionaries, because words have been included into the dictionaries as they appear in the text corpus. For example, for a verb different word forms can be derived from its stem form, if it carries the subject, the object, an auxiliary verb, a preposition, etc. The number of forms of the verb ጀመረ [dZAmArA] ’to begin’ is given in Table 5.1. Some of its forms are given in Table 5.2, page 72.

Dictionary size	No. of Forms of ጀመረ [dZAmArA]
28666	103
51489	167
114117	315

Table 5.1: Number of Forms of a Verb in Different Vocabulary Sizes

All the verbs have the potential for such a variation and most of these word inflections are

ጀመራችሁ [dZAmAratSIhu]	ጀ መ ራ ች ሁ [dZA mA ra tSI hu]	'You started'
ጀመረ [dZAmArA]	ጀ መ ረ [dZA mA rA]	'He started'
ጀመረች [dZAmArAtSI]	ጀ መ ረ ች [dZA mA rA tSI]	'She started'
ጀመርኩ [dZAmArIku]	ጀ መ ር ኩ [dZA mA rI ku]	'I started'
ጀመርን [dZAmArInI]	ጀ መ ር ን [dZA mA rI nI]	'We started'
ጀመሩ [dZAmAru]	ጀ መ ሩ [dZA mA ru]	'They started'
የጀመሩትን [jAdZAmArutInI]	የ ጀ መ ሩ ት ን [jA dZA mA ru tI nI]	'The one that they started'

Table 5.2: Different Amharic Verb Forms

realized in our dictionaries. In addition, nouns and pronouns also carry different elements of a sentence such as conjunctions, articles, possession marks and prepositions.

The dictionaries in the HTK system explicitly contain silence models as part of a pronunciation. The word "silence" is, therefore, included in our dictionaries and mapped to long silence in the speech at the beginning and end of a sentence. To model the short pause, all of the dictionaries add an "sp" at the end of every word pronunciation.

Currently, our pronunciation dictionaries do not handle the difference between geminated and non-geminated consonants, the variation of the pronunciation of the sixth order grapheme, with or without vowel, and the absence or presence of the glottal stop consonant.

Gemination of Amharic consonants range from a slight lengthening to much more than doubling. In the actual version of the dictionary, however, they are represented with the same transcription symbols. We have assumed, without conducting any experiments, that this inaccuracy in the transcription may be compensated by the looping state transitions of the HMM. We also expect that the geminated consonants and CV syllables containing them to have a higher loop probability than the non-geminated ones.

The sixth order vowel እ [I] is assumed to be pronounced in all the positions it can be used in Amharic writing, while it may not be pronounced in some cases. For example, the syllable ር [rI] in the word ጀመርን [dZAmArInI] 'we started' may be realized with, or without its vowel element. This problem is reflected in both the syllable- and phone-based pronunciation dictionaries.

In the syllable-based pronunciation dictionary, we have only one sequence of syllables for all of the words which contain the sixth order vowel that may or may not be uttered, and consequently should have different sequences of syllables (even different syllable structure). For example, we used the same symbol for the syllable ር [rI] in the word ጀመርን [dZAmArInI], whose vowel part may not be realized, and in the word በርዞ [bArIzo] 'he diluted with water' that is always realized with its vowel sound. That leads to force a syllable model to capture two different sounds: a sound of a consonant followed by a vowel and a sound of the consonant only.

The word ምግብ [mIgIbI] 'food', for instance, is mostly pronounced without the last vowel. But in our phone-based pronunciation dictionary, it is transcribed as if this vowel is always realized. That forces the model of the vowel to deal with a sound that is not of its nature.

A similar problem occurs with the glottal stop consonant እ [?] which may be realized or not. We have tried to solve this problem by using only the vowel that is next to this consonant for the phone-based pronunciation dictionary. Our decision to do so is based on an experiment using two different phone-based pronunciation dictionaries: one with the presence of the glottal stop in all the words where it occurs in the orthography, and another without it. For example, in the former dictionary the word በአል [bA?AlI] 'holiday' is transcribed as b A ? a l I while in the latter dictionary it is transcribed as b A a l I. The recognizer that used the latter dictionary performed better and, therefore, we generally omitted the glottal stop consonant. This method, however, can not be applied to the syllable-based pronunciation dictionary.

5.2.1 Syllable-based Pronunciation Dictionary

For the syllable-based pronunciation dictionary all 233 distinct CV syllables of Amharic are used. A sample of pronunciations in the canonical and alternative pronunciation dictionaries is given in Table 5.3, page 74. The alternative dictionary contains up to fifteen pronunciation variants per word form. Table 5.4 illustrates some cases of the variation. The encoding is given in Appendix C.

Words	Canonical Pronunciation	Alternative Pronunciation
CAmA	CA mA sp	CA mA sp Ca mA sp
Henedasu	He ne da su sp	He ne da su sp ne da su sp
HiteyoPeyA	Hi te yo Pe yA sp	Hi te yo Pe yA sp Hi te yo Pi yA sp Hi to Pe yA sp te yo Pe yA sp to Pe yA sp

Table 5.3: A Sample of Syllable-based Amharic Canonical and Alternative Pronunciation Dictionaries

Words	Number of pronunciation variants
HiteyoPeyAweyAne	15
HiteyoPeyAwiyAne	10
nawe	8
HiheHadige	8
HiheHadEge	8
yaHiteyoPeyAne	7
yaHiteyoPeyA	7
weseTe	7
miniseteru	7
HiteyoPeyAweyAnene	7
yaganezabe	6
HiteyoPeyAne	6
HiteyoPeyA	6
HegeziHabehEre	6
HegalexelAcehuAlahu	6
yehenene	5

Table 5.4: Number of Alternative Pronunciations of Some Words

5.2.2 Phone-based Pronunciation Dictionary

The overall number of phones used in the canonical phone-based pronunciation dictionary is forty-two, including the ”sil” and ”sp” units. We have stated in Chapter 2 that Amharic has thirty-one consonants and seven vowels which adds up to thirty-eight phones. As it can be seen in Table 2.1, there are only four rounding consonants. But we have realized that almost all Amharic consonants can be rounded.

In order to decide on how to deal with this difference between the rounded and the unrounded consonants, we have conducted an experiment using two pronunciation dictionaries. One of them transcribes this difference with different symbols while the other uses the same symbol, but their rounding feature is marked with a 'ue' symbol that is attached to the next vowel. The latter dictionary maps the rounded sound of the consonants and the next vowel to a symbol that represents a separate "rounding" vowel. Using these two pronunciation dictionaries, we have developed triphone-based recognizers and the recognizer that uses the latter version of our pronunciation dictionary achieved a better word recognition accuracy.

On the basis of the experiment, we have decided to model the rounding of all the consonants as separate "rounding" vowels that increased the number of our vowels to 12 (7 "normal" vowels and 5 "rounding" vowels). We have also modeled the consonant part of the syllable ኸ [hA] separately representing it by K, which could be mapped to the consonant (h). The forty-two phones in our pronunciation dictionary include, therefore, 28 consonants - 7 "normal" vowels, 5 "rounding" vowels, 2 silence models (sil and sp).

Due to time constraints, no phone-based alternative pronunciation dictionary has been prepared. A sample of phone-based pronunciations is given in Table 5.5. The encoding is given in Appendix C.

Words	Pronunciation
CAmA	C A m A sp
CesaNenate	C e s a N e n a t e sp
CoKa	C o K a sp

Table 5.5: A Sample of Phone-based Amharic Canonical Pronunciation Dictionaries

5.3 The Language Model

As we can see from equation 3.7, page 42, the language model is one of the most important knowledge sources for a large vocabulary speaker independent recognition system. It is also indicated in Figure 3.1 as one of the main components for the development of an ASRS. The language model incorporates knowledge of the language, such as its syntactic and semantic information, in the ASRS by providing the probabilities that a word or string of words is/are followed by another word in a given text.

As no usable language model for Amharic does exist, we had to develop one for our experiment. Therefore, we have trained bigram language models using the HTK statistical language model development modules. Had enough training text been accessible to us, we would have developed trigram language models.

Training a statistical language model requires text data that consist of millions of words (Jelinek 1990). But we have a training text consisting of less than 750,000 words. Due to this shortage of training text we have included sentences from the other test sets for developing language model for a test set (5000 development test set e.g.). For example, when we develop the language model for the 5,000 development test set, we used sentences from the other test sets. The sentences in this test set could also be used to train the language model for the other test sets. In this way, we have had training texts of approximately 75,000 sentences and 900,000 words, including the 10,875 sentences and 104,054 words recorded for the training speech data. The same language models are used to evaluate both syllable-based and phone-based recognizers of the same vocabulary size. The perplexities of our language models are given in Table 5.6, page 77.

The high perplexity of the language models may be attributed to the morphological property of Amharic that keeps the frequency of words very low as indicated in Table 5.7, page 77.

To have a comparative view of this property of the language, we took an English text that has 13,126 different words, as an example, and Amharic text of 24,284 different words

Test sets	Vocabulary size	Perplexity
Development	5000	105.39
	20000	167.889
Evaluation	5000	101.589
	20000	165.519

Table 5.6: Perplexities of the Language Models

Frequency	No of words
1000 and more	61
100-999	1086
1	59419(52%)

Table 5.7: Frequency of Amharic Words

and show their word frequency in Table 5.8, page 77. The English text is taken from WSJCAM0.

Frequency	Amharic		English	
	No of words	Percent	Percent	No of words
1000 and more	1	0.004	0.40	52
100-999	48	0.2	4.2	550
10-99	1,456	6.0	25.6	3,362
2-9	6,233	25.7	62.7	8,227
1	16,547	68.1	7.1	932

Table 5.8: Frequency of Words in Amharic and English Texts

The Table shows that only 0.004% of Amharic words occurred 1000 and more times while 0.4% of English words occurred 1000 and more times in the text. On the contrary, 68.1% of Amharic words occurred only once, while only 7.1% of English words are rare.

5.4 Speech Feature Extraction

During the process of feature extraction, the speech signal is converted into a stream of feature vectors which contain the information about a given utterance that is important for its recognition. Parameterization is performed not only for size reduction of original speech signal data, but also for pre-processing of that signal into a form fitting requirement of the following 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. They are derived by a mathematical transformation which computes the inverse Fourier transform of the log-spectrum of the speech signal. We have preferred to use MFCCs for our experiment of developing Amharic ASRSs.

We have performed the following functions using the HTK feature extraction module by setting the required configuration for the HMM-based classifiers:

1. The speech signal is divided into frames of size 25ms with a frame rate of 10ms. Attempts to use a larger frame size and frame rate for our syllable-based recognizers resulted in performance degradation - although it increased the speed of the recognition. However, we did not have the time to do experiments with more than two frame sizes (50ms and 100ms) and frame rates (20ms and 50ms).
2. MFCC features are extracted based on the short-term Fourier spectrum and the power or magnitude of Fourier spectrum was computed for every speech segment.
3. Due to the shortage of training speech, only a variance vector is used instead of a full co-variance matrix.
4. Feature vectors are completed by delta and acceleration coefficients.

We have adopted the configuration parameters of feature extraction that specify the above-mentioned points, and other information for the feature extraction module of the toolkit from HTK recommendations (Young et al. 2002). They are given in Appendix B.

5.5 Syllable-based Acoustic Model

The syllable-based acoustic model is trained for all of the 233 Amharic syllables using the feature extracted as explained in Section 5.4. The training is performed using the embedded re-estimation tools of the HTK. We have conducted different experiments with various sizes and topologies of HMM, and selected the best performing ones. To check model quality between two training cycles, the development test sets of 5000 and 20000 vocabulary are used. Finally, the well-performing recognizers on the development test sets are evaluated using the 5000 and 20000 vocabulary evaluation test sets.

5.5.1 Training

Training of an HMM for speech recognition is the refinement of a set of λ's (π, A, B) that best represent the training speech. Where: (π) denotes the initial state distribution; (A) denotes the state-transition probability distribution; and (B) denotes the observation symbol probability distribution.

The training commences with initial values for λ's and refines them based on the forward and backward algorithm called the Baum-Welch algorithm. The initial values of $(\pi), (A)$ and (B) are determined as a part of initialization of the model.

We have stated in Section 3.5.2, that there are two methods of initializing a set of HMMs using the HTK toolkit: bootstrapping and flat start.

Both of the initialization methods have been investigated experimentally for syllable HMMs. We have initialized HMMs with both methods and trained them in the same way. The HMMs that have been initialized with the flat start method performed better (40% word recognition accuracy) on development test set of 5,000. We have, therefore, continued our subsequent experiments with the flat start initialization method.

The problem with the bootstrapping approach is that any error of the labeler strongly affects the performance of the resulting model because consecutive training steps are influenced by the initial value of the model. That may be why we did not benefit from the use of our segmented speech, which has been transcribed with a speech recognizer that has a low word recognition accuracy, and edited by non-linguist listeners. A lack of time prevented us from exploring the problem of the bootstrapping approach of initialization in more detail.

The initialized models are re-trained with the Baum-Welch re-estimation procedure described earlier in Chapter 3. With the Baum-Welch reestimation procedure, we start with an arbitrary model and iterate to improve it. All we need for the start of this process are sub-word-level transcriptions of the utterances, even without alignment to the speech signal. The Baum-Welch algorithm takes the initial parameters and computes the training data's likelihood. It then adjusts the probabilities in such a way that the model guarantees that the next iteration will produce a better likelihood. Although this procedure does not guarantee us to find a globally optimal solution, it usually produces satisfactory estimates (Young et. al 2002).

After the second iteration of the re-estimation of syllable models, short-pause models are added and the silence model is extended by allowing individual states to absorb the various impulsive noises in the training data. This is achieved by adding extra transitions from the first emitting state to the last emitting state in the silence model and tying the emitting state of a one-state short-pause (sp) model to the center state of the silence model.

5.5.2 HMM Topologies

To our knowledge, there is no topology of HMM model that can be taken as a rule of thumb for modeling syllable HMMs, especially, for Amharic CV syllables. To have a good HMM model for Amharic CV syllables as recognition units, one needs to conduct an experiment to select the optimal model topology. Designing an HMM topology consists of choosing an appropriate number of states, the allowed initial states and the allowed transitions. That has to be done with proper consideration of the size of the unit of recognition and the amount of the training speech data. This is due to the fact that as the size of the recognition unit increases and the size of the model (in terms of the number of states and number of transitions) grows, the model requires more training data.

For our syllable model, we, therefore, carried out a series of experiments using a left-to-right topology with and without jumps and skips, with a different number of emitting states (3, 5, 6, 7, 8, 9, 10 and 11) and different number of Gaussian mixtures (from 2 to 98).

By jump we mean skips from the first non-emitting state to the middle state and/or from the middle state to the last non-emitting state. We have used this topology to develop a solution for the problem of the irregularities in the realization of the sixth order vowel እ [I] and the glottal stop እ. We conducted an experiment using HMMs with a jump from the middle state to the last non-emitting state for all of the CV syllables with the sixth order vowel እ [I], and a jump from the first emitting state to the middle state for all of the CV syllables with the glottal stop እ [?]. These topologies have been chosen so that the models recognize the absence of the vowel and the glottal stop consonant of CV syllables. When we analyze the trained models of the above-mentioned syllables, we can observe that they favor such a jump. This confirms that these phones of the syllables are not always uttered. This situation is shown in Figures 5.1 - 5.3. We have used the model of the syllable ይ [jI] to exemplify the models of all the syllables with the sixth order vowel.

The transition probabilities in Figure 5.1 show that the models of the glottal stop consonant with the sixth order vowel tend to start emitting with the 4^{th} state with a probability of 0.72. The model also has accumulated a considerable probability (0.38) to jump from the 4^{th} state to the last (non-emitting) state.

The model of this consonant with the other vowels (our example is the 5^{th} order vowel and shown by Figure 5.2) tend to start emitting with the 4^{th} state with a probability of 0.68. This is two times the probability (0.32) of its transition to the 2^{nd} state.

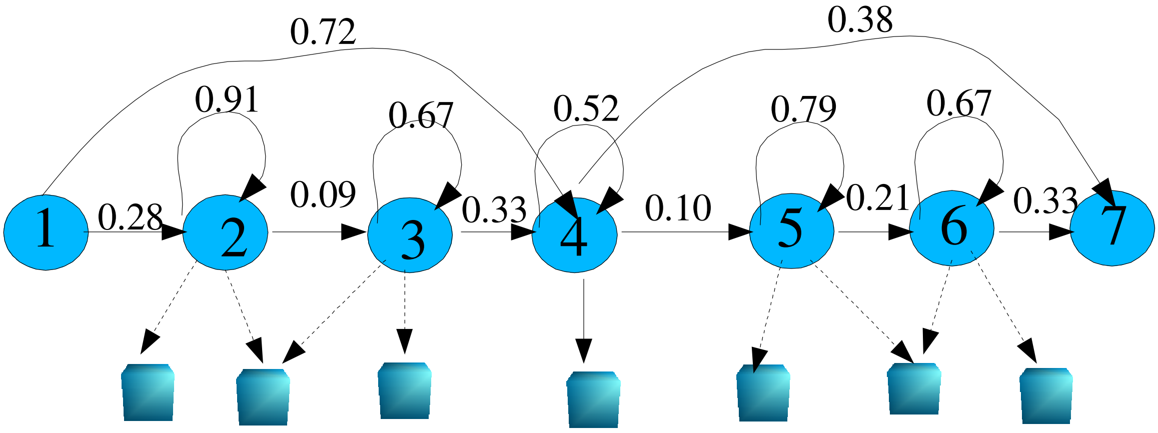

Figure 5.1: An HMM Model of the Syllable እ [?I]

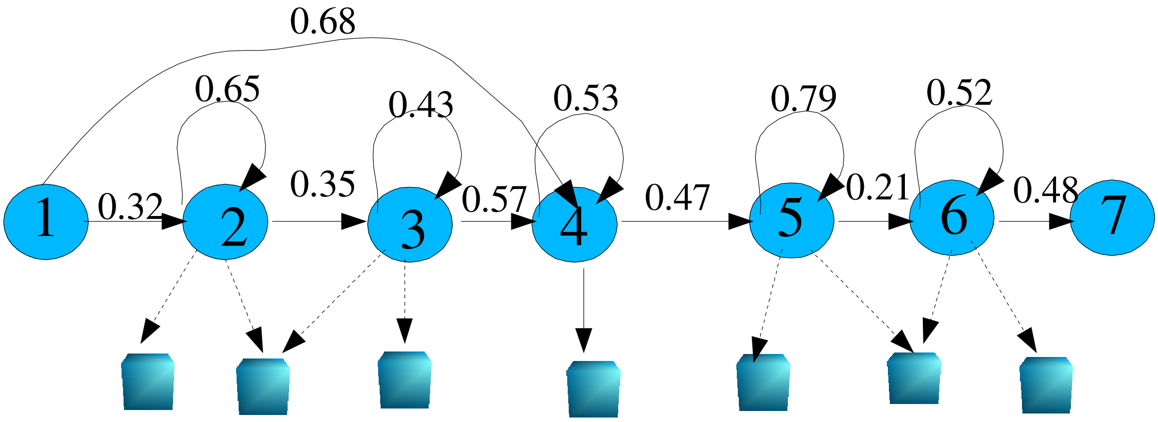

Figure 5.2: An HMM Model of the Syllable ኤ [?e]

The models of the other consonants with the sixth order vowel, which are exemplified by the model of the syllable ይ [jI] and shown by Figure 5.3, tend to jump from the 4^{th} state to the last (non-emitting) state with a probability of 0.39, which is considerably greater than that for continuing with the next state (0.09).

For models with skips, we have limited the number of states to be skipped to one because the amount of training speech that we have is too limited to train the additional transition probabilities for skipping two states.

To determine the optimal number of Gaussian mixtures for the syllable models, we have

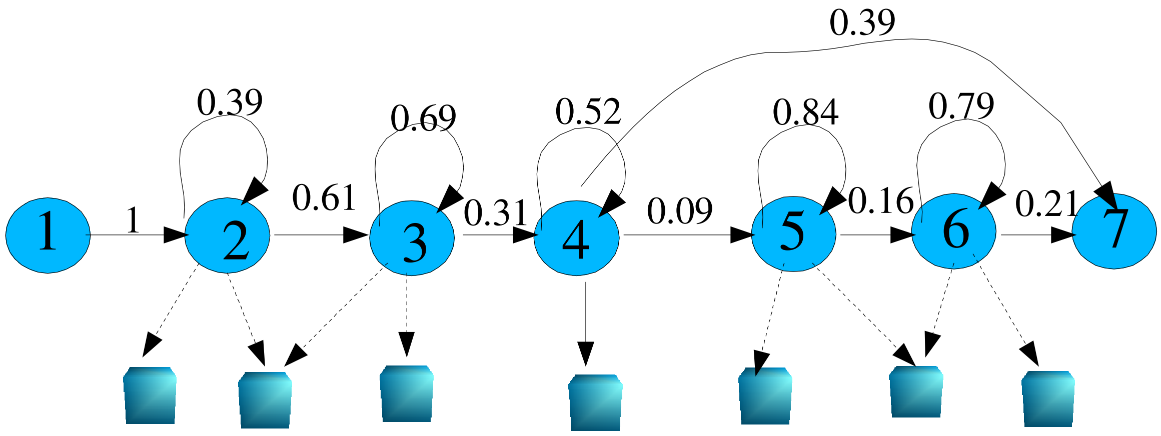

Figure 5.3: An HMM Model of the Syllable ይ [jI]

conducted a series of experiments by adding two Gaussian mixtures for all of the models until the performance of the model starts to degrade. Considering the difference in the frequency of the CV syllables, we have tried to use a hybrid number of Gaussian mixtures. By hybrid, we mean that Gaussian mixtures are assigned to different syllables based on their frequency. For example: the frequent syllables, like ን [nI], are assigned up to fifty-eight while rare syllables, like ጲ [p'i], are assigned not more than two Gaussian mixtures.

5.5.3 Recognition results

We present recognition results of only those recognizers which have competitive performance to the best performing model. For example: the performance of the model with 11 emitting states with skips and hybrid Gaussian mixtures is more competitive than those with 7, 8, 9, and 10 emitting states. We have also systematically left out test results which are worse than those presented in Table 5.9. For example, it is obvious that word recognition accuracy of a recognizer is better on a 5,000 test set than on a 20,000 test set; likewise, the accuracy of a recognizer using only the acoustic models will be lower than its performance when combined with a language model. Accordingly, we have included in Table 5.9 only relevant results.

From Table 5.9, we can see that the model with five emitting states, with twelve Gaussian mixtures, without skips and jumps has the best word recognition accuracy.

A CV syllable consists of two phones, a consonant and a vowel. When we consider the most commonly used number of HMM states for phone-based speech recognizers (three emitting states), we expect a model of six emitting states to be best. But the result of our experiment shows that a CV syllable-based recognizer performed better with only five emitting states compared to all the other topologies we have used. This may be attributed to the fact that speech parameters of a consonant, a vowel and the context between them are used to train one HMM model.

The fact that the model of five emitting states need no skips lets us conclude that CV syllables have consistently more than five speech frames.

No. of emitting states	Transition Topology	No. of Mix	Models					
			Acoustic Model(AM)		Acoustic Model+ Language Model(LM)		AM + LM + S. Adaptation	
			5k	20k	5k	20k	5k	20k
3	without skip and jump	18	62.85		88.82	85.73		
		hybrid	60.87		87.63		88.50	
	with skip	12			69.2			
	with jump	12	43.74		79.94			
5	without skip and jump	12	69.29		88.99		**89.80**	87.69
		hybrid	60.04					
	with skip	12			85.77			
	with jump	12	54.53		84.60			
6	without skip and jump	10	66.73		86.99			
		hybrid	65.58		86.40		86.58	
	with skip	hybrid	53.77					
	with jump	8	65.68		87.58			
11	with skip	12	55.04					
		hybrid	71.83		**89.21**	87.15	89.04	

Table 5.9: Recognition Accuracy of Syllable Models on the Development Test Sets

As we can see from Table 5.9, models with three emitting states do have a competitive performance with 18 (with word recognition accuracy of 88.82) and hybrid (with word recognition accuracy of 88.50) Gaussian mixtures. They have the least number of states of all the models that we have developed. Nevertheless, they require more storage space

(33MB with 18 Gaussian mixtures and 34MB with hybrid Gaussian mixtures) than the best performing syllable models (32MB). They also have a larger number of total Gaussian mixtures[1] (30,401 with 18 Gaussian mixtures and 31,384 with hybrid Gaussian mixtures) than the best performing syllable model that uses a total of 13,626 Gaussian mixtures.

The other model topology that is competitive in word recognition performance is the model with eleven emitting states, with skip and hybrid Gaussian mixtures, which has a word recognition accuracy of 89.21%. It requires the biggest memory space (40MB) and uses the largest number of total Gaussian mixtures (36,619) of all the syllable models we have developed.

We have evaluated the top two models with regard to their word recognition accuracy on the evaluation test sets. Their performance is presented in Table 5.10. As it can be seen from the Table, the model with the better performance on the development test sets also showed the same results with the evaluation test sets. We can, therefore, say that the model with five emitting states without skips and twelve Gaussian mixtures is preferable not only with regard to its word recognition accuracy, but also with regard to its memory requirements. Figure 5.4, page 86 shows the topology of the best performing syllable model.

No. of emitting states	No. of Mix	Models					
		Acoustic Model(AM)		Acoustic Model+ Language Model(LM)		AM + LM + S. Adaptation	
		5k	20k	5k	20k	5k	20k
5	12					90.43	87.26
11	hybrid			89.36	87.13		

Table 5.10: Recognition Accuracy of Syllable Models on the Evaluation Test Sets

[1] We counted the Gaussian mixtures that are physically saved, instead of what should actually be. For example, we expect $5 \times 12 \times 234 = 14{,}040$ total Gaussian mixtures of a recognizer that has 5 emitting states, 12 Gaussian mixtures and 234 HMMs. However, HTK saves only 13,626 of these Gaussian mixtures that are properly trained by the available training data.

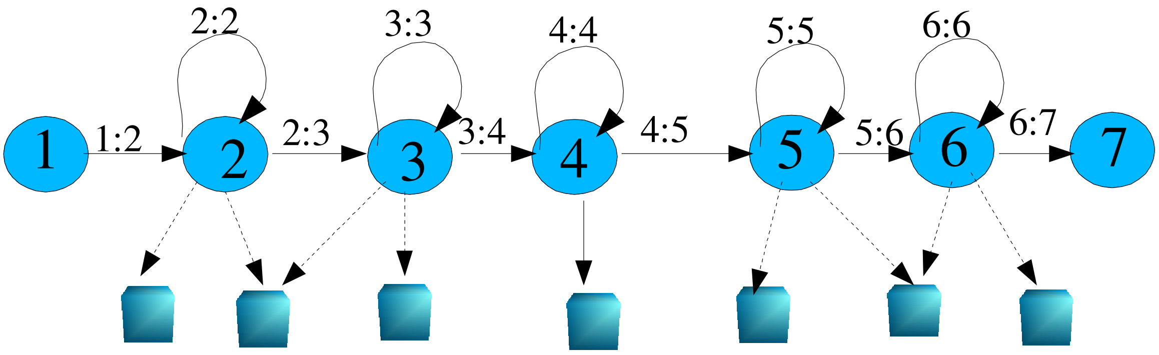

Figure 5.4: The Best Performing Syllable HMM Model

5.6 Phone-based Acoustic Model

For the phone-based systems an HMM with a 3-state left-to-right topology was used. To decide on the use of the model with or without skips, we have developed recognizers using both topologies with different Gaussian mixtures. We have also conducted experiments on models with and without jumps from the first emitting state to the last non-emitting state.

We initialized the model with the flat start method and used it for all the monophones. The monophone models are then retrained, short-pause models are added and the silence model is extended slightly, as it has been done for the silence model in the syllable-based recognizers. This set of monophone models has been retrained two times as recommended in the HTK book (Young et. al 2002).

In a next step triphone models have been derived by cloning the respective monophone models. Cloning is the process of duplicating a monophone model for all of the triphones that represent the different contexts of the monophone. For example, at the beginning all the triphones that have (b) as their center, such as (a-b+c) and (c-b+d), have the same model, which is the facsimile of the monophone model (b).

5.6.1 Triphone HMM

We have created our inter-word and context-dependent triphone HMMs in two steps. Firstly, the monophone transcriptions are converted into word internal triphone transcriptions, and a set of triphone models is created by cloning the monophones and then re-estimating using triphone transcriptions.

Secondly, similar acoustic states of these triphones are tied using decision-tree based state-clustering, to ensure that all state distributions can be robustly estimated. Tying is necessary since some triphones occur only once or twice and would be poorly estimated if tying would have not been done.

Tying could negatively affect the performance if done indiscriminately. Hence, it is important to only tie parameters which have little effect on classification. These are the transition parameters which do not vary significantly with acoustic context, but nevertheless need to be estimated accurately. Therefore, we have tried to tie only those parameters of our triphones which have little effect on their classification. To this end, we used the HHED module of the HTK that uses decision trees to clustered states and then tie each cluster. The use of decision trees is based on asking questions about the left and right contexts of each triphone. The decision tree attempts to find those contexts which make the greater difference to the acoustics and which should, therefore, distinguish the clusters.

Once the context-dependent models have been cloned and tied, the new triphone set has been re-estimated using a tied triphone list and the triphone transcriptions. Finally we have augmented the context-dependent triphone models with the context-independent (sp and sil) models to get an extended list.

Using the above procedure of developing triphone models, we have conducted experiments with different HMM topologies such as HMMs with and without skips and a different number of Gaussian mixtures.

To deal with the irregular realization of the vowel እ [I] and the glottal stop እ [?], we also conducted an experiment using a jump from the first emitting state to the last non-emitting state for these phones. The topology of the model is shown in Figure 5.5. The transition probabilities of the models of the vowel and of the glottal stop consonant are given in Tables 5.11 and 5.12, respectively.

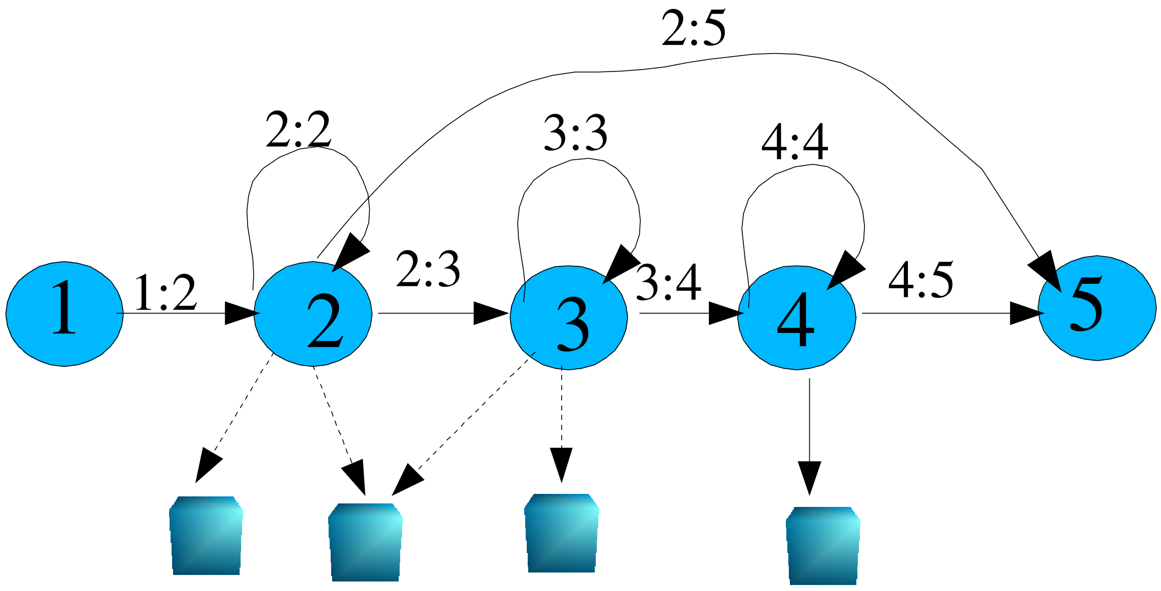

Figure 5.5: An HMM Model of the Vowel እ [I]

	1	2	3	4	5
1	0	1	0	0	0
2	0	0.49	0.15	0	0.36
3	0	0	0.63	0.37	0
4	0	0	0	0.68	0.32
5	0	0	0	0	0

Table 5.11: Transition Probabilities of the Vowel እ [I]

The transition probabilities in Tables 5.11 and 5.12 show that the models tend to jump from the first emitting state to the last (non-emitting state) with probabilities of 0.36 and 0.29, respectively. This indicates that the Amharic vowel እ [I] and the glottal-stop consonant እ [?] are not uttered in a number of the places where they must be written.

	1	2	3	4	5
1	0	1	0	0	0
2	0	0.59	0.12	0	0.29
3	0	0	0.69	0.31	0
4	0	0	0	0.57	0.43
5	0	0	0	0	0

Table 5.12: Transition Probabilities of the Glottal-stop Consonant እ [?]

5.6.2 Recognition results

As it is presented in Table 5.13, page 89, the set of models with skip, without jump and twelve Gaussian mixtures has the best performance among all of the different topologies. The topology of best performing set of models is shown in Figure 5.6, page 90.

Vocabulary size	Transition Topology	No. of Mixture	Models		
			Acoustic Model(AM)	AM+ Language Model(LM)	AM + LM + S. Adaptation
5,000	without skip and jump	20	73.86	90.46	88.67
	with skip	12	73.49	90.94	90.85
	with jump	20	63.27		
20,000	without skip and jump	20		83.80	

Table 5.13: Recognition Accuracy of Triphone Models on the Development Test Sets

We have presented the recognition results for acoustic models with a zero-gram language model, a bigram language model without speaker adaptation, and a bigram language model with speaker adaptation. The indicated Gaussian mixtures are only optimal for the respective model, i.e. the model with skips has the best performance with twelve Gaussian mixtures while the model with jumps performs best with twenty Gaussian mixtures.

When we compare the performance of the acoustic models, with the zero-gram language model and without speaker adaptation, we see that the triphone model with no skips and jumps performs best (with 73.86% word recognition accuracy) with twenty Gaussian

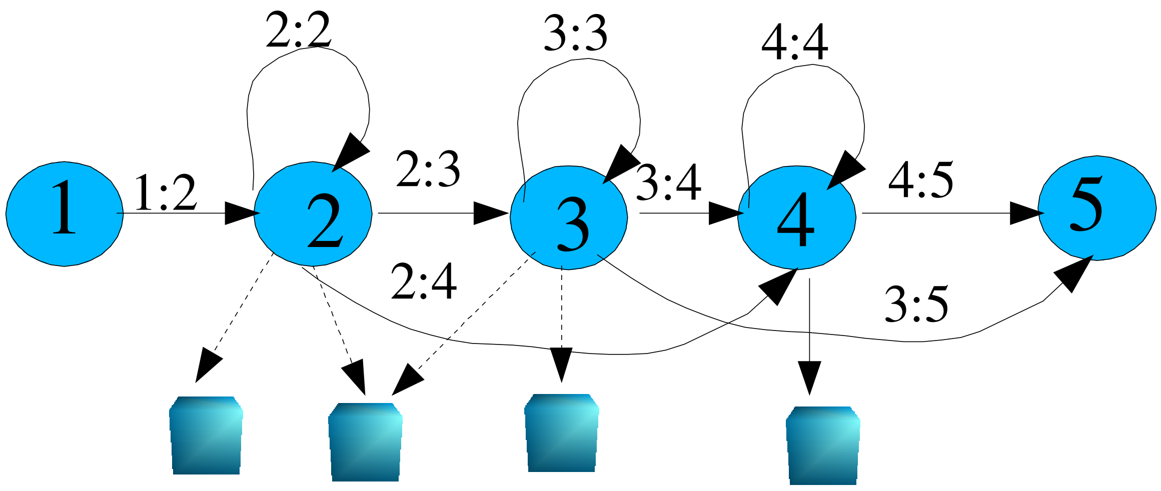

Figure 5.6: The Best Performing Triphone HMM Model

mixtures. Using only twelve Gaussian mixtures, the model without skips and jumps has 73.49% word recognition accuracy. These are, therefore, the two competitive triphone models which we continued to test their performance with the use of language model and speaker adaptation.

With the use of the bigram language model and speaker adaptation, the model with skips performs better than the model without skips.

We have also evaluated these two triphone models on the evaluation test sets. The result of their word recognition accuracy is presented in Table 5.14, page 91. Their performance on the evaluation test sets show that the model with skips has better word recognition accuracy.

Considering the size of the models, the ones without skips use twenty Gaussian mixtures resulting in 51,625 physically saved Gaussian mixtures and a memory requirement of 57MB. On the contrary, the models with skips use only twelve Gaussian mixtures and, consequently, have a total of only 33,702 physically saved Gaussian mixtures that require a memory of only 38MB.

Therefore, it is clear that the triphone model with skips is preferable with respect to its

Vocabulary size	Transition Topology	No. of Mixture	Models		
			Acoustic Model(AM)	AM+ Language Model(LM)	AM + LM + S. Adaptation
5,000	without skip and jump	20		90.31	
	with skip	12		91.31	
20,000	without skip and jump	20		88.40	
	with skip	12		89.75	

Table 5.14: Recognition Accuracy of Triphone Models on the Evaluation Test Sets

word recognition accuracy, total physical size in terms of the number of Gaussian mixtures, and memory requirement.

5.7 Limiting Factors of the Recognizers' Performance

Here, we are not going to list all of the factors that limit the performance of ASRSs. It is our intention to only point out some factors which affected the performance of the recognizers that we have developed for Amharic. More emphasis is given to factors which affect the intended comparison between syllable- and phone-based Amharic ASRSs.

The inflectional property of Amharic leads to the effect that many words in the pronunciation dictionary share the same sequence of sub-word units. Therefore, many words vary only slightly. Consequently, the number of competing sequences of acoustic models (HMMs) to be considered during recognition increases.

This property of the language affects the language model, in that the frequency of the majority (more than 90%) of Amharic words is below ten. With such a training text, it is obvious that our statistical language model is not well trained. The perplexities of our language models are also relatively high.

Moreover, as already discussed in Section 5.2, our pronunciation dictionary does not handle the irregular realization of the sixth order vowel and the glottal stop consonant. This problem leads to a mismatch between the training speech and the sub-word transcription of the training text and, subsequently, to a reduction of the quality of the HMM models for these phones and the syllables containing these phones.

The above two factors affect both systems, while all of the following factors introduce biases in favor of phone-based recognizers.

Phone-based recognizers were able to benefit from using a pronunciation dictionary, which ignores the glottal stop consonant in the transcription of words that contain it while the syllable-based recognizers could not. The phone-based recognizers that are developed using this pronunciation dictionary have better word recognition accuracy than those which use a pronunciation dictionary that transcribes words as they are written. We were not able to develop a syllable-based pronunciation dictionary that ignores the glottal stop consonant, because ignoring a consonant in the CV sequence will leave the vowel to stand alone and leads us to develop a hybrid recognizer that uses CV syllables and phones together, which goes beyond the scope of this thesis.

The other limiting factors relate to the shortage of training speech data. During training, HTK removes mixtures whose weight falls below a certain threshold (1.0e-05, the default value of MINMIX). This leads to a mis-match between the expected number of mixture components and the actual number of mixture components in the HMM file and a reduction of the number of mixture components in one of the terminal nodes of the regression tree, which is generated for the purpose of speaker adaptation. Let us give two examples that show the case.

- In a triphone-based model with skips and twelve Gaussian mixtures, the components are reduced from 2,840 (expected) to 1,418 (the actual number). The expectation is with regard to the number of physical triphone models, not of the logical models that are tied to the physical models.
- In a syllable-based model with five emitting states, without skips and jumps and with twelve Gaussian mixtures, the components are reduced from 809 (expected) to 395

(the actual number).

We could also see this problem in the difference between the expected total number of Gaussian mixture, and the number of the Gaussian mixture that are physically saved in the HMM files, as given below:

- In the triphone model of three emitting states and twelve Gaussian mixtures that contains a set of 4,099 physical models, we expect $3 \times 12 \times 4,099 = 147,564$ Gaussian mixtures. Actually, we have only 33,702 Gaussian mixtures that are physically saved.
- In the case of the syllable model of five emitting states and twelve Gaussian mixtures that contains 234 physical models, we expect $5 \times 12 \times 234 = 14,040$ Gaussian mixtures. But we have only 13,626 physically saved Gaussian mixtures.

For the phone-based recognizer, the shortage of training speech data could partly be compensated by tying the components of the triphones. This enables many triphone models to share training speech data, thereby reducing the number of triphone models from 5,092 logical models to 4,099 physical models.

On the other hand, syllable-based recognizers are not tied at all and are more seriously affected by a shortage of training speech. Tying, as noted in section 3.5.2, is the well-known solution to the shortage of training speech data in the development of ASRSs.

A triphone model is a model of that phone in one context, and another triphone model used to train the speech of that same phone in another context. For example, (l-a+m) is a model of the vowel (a) in the context of consonant (l) on the left and (m) on the right. It uses another model when one of these contexts change. For example, (b-a+m) and (l-a+s) are different models of the same vowel.

On the other hand, the model of a CV syllable considers only one context of a phone, i.e. there are different models for a consonant in the context of different vowels. For example, (la) is a model of consonant (l) and vowel (a), while (lu) is another model of the same consonant and another vowel. Similarly, (ma) is a model of the consonant (m) and vowel (a), while (la) is another model of the same vowel and another consonant. In contrast to the triphone models, the CV syllable models consider only right context for a consonant and left context for a vowel. Therefore, the CV syllable models have access to less context information than the triphone models.

With these factors in mind, let us compare the syllable- and phone-based Amharic speech recognizers.

5.8 Comparison of the Syllable- and Phone-based Amharic ASRSs

Test Sets	HMMs	No. of Mix	Models					
			Acoustic Model(AM)		Acoustic Model+ Language Model(LM)		AM + LM + S. Adaptation	
			5k	20k	5k	20k	5k	20k
Development	Syllable with 5 emitting states	12	69.29		88.99		**89.80**	87.69
	Triphone with skip	12	73.49		**90.94**			
Evaluation	Syllable with 5 emitting states	12					**90.43**	**87.26**
	Triphone with skip	12			**91.31**	**89.75**		

Table 5.15: Recognition Accuracy of Triphone and Syllable Models on the Development Test Sets

From the point of view of their word recognition accuracy, Table 5.15 shows that the triphone model performed better than the syllable model. But we have to take note that the syllable model is not tied at all and may gain more from tying or use of more training speech data than the triphone model. To see this fact, let us compare our results with Kinfe's (2002) research that is conducted using training speech of only 170 vocabulary and fifteen training speakers.

Even though the CV-syllable was an attractive sub-word unit for Amharic, given

> the nature of language, it has resulted in a relatively poor performance, i.e., 84% on the training data and 70% on the testing data using the researcher's own voice. The same voice has resulted in 92% percent recognition accuracy for the phoneme-based recognizer on both training and testing data, while 94% and 90% recognition accuracies have been obtained for tied-state triphones on the training data and testing data, respectively.

In Kinfe's work, there is a difference of 20% word recognition accuracy between the triphone- and syllable-based recognizers, in favor of the triphone-based one. But in our work, which uses training speech of 28,666 vocabulary and 100 training speakers, the performance difference is less than 1% word recognition accuracy.

With regard to their requirement of storage space and recognition speed, the models of the best performing triphone-based recognizer uses 38MB space, while the syllable-based recognizer uses only 15MB space. This is because the syllable model has only a total of 13,626 Gaussian mixtures, while the triphone model has a total of 33,702 physically saved Gaussian mixtures.

Considering the speed of processing, the triphone-based model recognized ten sentences in thirty-eight minutes with 96.52% word recognition accuracy, while the syllable-based recognizer finished the same files in twenty-four minutes with 97.39% word recognition accuracy. The speed of the syllable-based recognizers is attributed to the smaller size of the set of acoustic models compared to the set of triphone-based acoustic models. The smaller the size of a set of acoustic models, in terms of the number of Gaussian mixtures, the lower the effort for computing the emission probability in the recognition process.

Chapter 6

Conclusions and Recommendations

6.1 Introduction

In the previous Chapters we have presented an overview of the Amharic language, and reviewed the theoretical basis of ASRSs development. As a required part of ASRSs development, we have prepared an Amharic speech corpus and presented the process of its preparation along with a brief description. Our main achievement has been the experimental result of our work, as presented in Chapter 5. In this Chapter we present our conclusive remarks and recommend future work that is to be done in the area of Amharic ASRSs development.

6.2 Conclusions

We have achieved the main objective of our research project which is the exploration of possibilities for the development of a large vocabulary, speaker independent and continuous speech recognition system for Amharic. The corpus that we have developed and used, as presented in Chapter 4, is a medium size Amharic read speech corpus. It can be used to develop a large vocabulary ASRS for Amharic. We have used the corpus to develop different syllable- and triphone-based ASRSs for Amharic which have word recognition accuracy of above 90%.

Therefore, the potential of developing Amharic ASRSs using the HTK that applies the HMM frame-work has been demonstrated by our work. We have also shown that it is possible to use the CV syllable as a basic unit of recognition to develop an Amharic ASRS. Exploration was done on different HMM topologies for Amharic CV syllables and phones.

From our experiments we have found that the optimal HMM topology for Amharic CV syllable is a left-to-right model with five emitting states, without skips and jumps and with twelve Gaussian mixtures. We have also discovered that a model of three emitting states, with skips and twelve Gaussian mixtures, is the topology of preference in triphone modeling for Amharic among those tested model topologies.

Along with checking our working hypothesis that is stated in Chapter 1, we have compared syllable-based and triphone-based models using different points of views, like word recognition accuracy, recognition speed and memory requirement of the models.

Based on the findings of our experiment, we claim that our working hypothesis is an acceptable hypothesis, because a syllable-based recognizer is competitive with the triphone-based recognizer (90.43% vs 91.31%) with regard to its word recognition accuracy. It is also faster and requires less memory space than the triphone-based recognizer. Furthermore, we expect more improvement in the performance of syllable-based recognizers compared to the triphone-based ones, if they are properly tied.

Therefore, in conclusion, we believe that the use of CV syllables is a promising alternative in the development of ASRSs for Amharic.

6.3 Recommendations for Future Works

As pointed out in the conclusion of our work, we have achieved considerable progress in the development of speaker independent LVCSRSs for Amharic. However, we see that there is a lot of work remaining in the area. In this Section, we would like to forward some of the areas that need researchers' attention in four parts: Speech corpora, language modeling, development of the acoustic models of speech recognizers, and the application of speech recognizers.

6.3.1 Speech Corpora

We have developed only a read speech corpus out of a number of other types of speech corpora that have been mentioned in Chapter 4. There remains a need for developing other types of corpora, like a spontaneous speech corpus, and making them available to researchers and developers of different Amharic ASRSs. We make no claim that complete work has been done in the area of developing a read speech corpus for Amharic. There is a need to add more training speech, as well as, improve our pronunciation dictionaries.

With regard to the amount of the training speech data, our corpus is only a medium size speech corpus of 20 hours of speech. When compared to the other speech corpora that contain hundreds of hours of speech data for training, e.g. the British National Corpus (1,500 hours of speech), CGN (Spoken Dutch Corpus) (total of 800 hours of speech and 104 hours of read-speech), CSJ (Corpus of Spontaneous Japanese) (650 hours of speech) and BREF-120 (A large corpus of French read speech) (100 hours of speech), then ours is only a starting point providing a minimum of the required speech data for Amharic. We highly recommend, therefore, an improvement of the corpus by recording and transcribing more training speech data.

Our pronunciation dictionary does not handle the problem of gemination of consonants and irregular realization of the sixth order vowel and the glottal stop consonant, which has a direct effect on the quality of the sub-word transcriptions. Proper editing of the pronunciation dictionaries which, however, requires a considerable amount of work, certainly will result in a higher quality of sub-word transcription for the corpus, and consequently in the improvement of the recognizers' performance.

6.3.2 Language Model

We have pointed out that the inflectional property of Amharic increased the perplexity of our language models. In such a language one needs to develop a sub-word language model (Kim 1997), or use a huge training text corpus to develop a good statistical language model (Jelinek 1990). Both possibilities are research directions that we would like to recommend in this area of Amharic speech recognition.

6.3.3 Acoustic Models

Although the acoustic models that we have developed have word recognition performance of more than 70%, a further performance improvement can be expected. To this end, new methods for initializing the models, a refinement of the models by the use of tying, and speaker adaptation techniques need to be investigated.

The use of segmented speech data for initialization is a proven method in the community of ASR, but any segmentation error also has a significant negative impact on the performance of the models. Due to limitations of time, human and financial resources, we could not produce an error free segmentation of our speech at syllable and phone levels, but expect an improved performance of Amharic speech recognizer as a result of better initial parameters.

We could not work on tying the parameters of the syllable models. Since tying is one way of minimizing the problem of training speech shortage, tying the syllable models would possibly result in a gain of some degree of performance improvement. The CV syllable model is a model of a consonant and a vowel. We have seven different CV syllable models (11 for some consonants that have rounding feature) that have the same consonant but different vowels and thirty-one different CV syllable models that have the same vowel but different consonants. The right half of different models that are composed of one vowel but different consonants can be tied together. In the same understanding, the left half of different models that are made of one consonant but different vowels can be tied together.

We have also seen that the CV syllables models have less context information than the triphone models. This can be addressed by exploring methods of modeling contexts between syllables. The development of a bi-syllable model together with the proper tying is recommended, to gain a considerable performance improvement for the syllable-based recognizers.

When we had assigned the same number of Gaussian mixtures to all of the syllables, less frequent syllables could not train a large number of mixtures components. As a solution, we have used a different number of Gaussian mixtures according to the frequency of the syllables in the corpus, which we called a hybrid mixture increment. Nevertheless we do not claim to have really used the optimal number of Gaussian mixture for each of the syllables. Therefore, this is one area of work that requires further investigations.

Our models could not gain the performance improvement that is expected from speaker adaptation. We have used only maximum likelihood linear regression (MLLR) technique, and have applied only the procedure that is described in the HTK tutorial. But, the MLLR technique itself is flexible and can be used in different ways, like varying the number of transformations to fit the available adaptation speech data (Young et.al. 2002). Furthermore, the maximum a posteriori (MAP)(Young et.al. 2002) and other adaptation approaches (Woodland 2001) and (Chen 1997), for example, have not been investigated. We, therefore, recommend working on the application of a speaker adaptation technique that properly utilizes the available speaker adaptation speech data and gains a significant improvement in the performance of the recognizers.

6.3.4 Application of the Speech Recognizers

The development of an ASRS is not an end by its own. There remains much work in the area of incorporating the Amharic ASRS in different areas that benefit from the use of ASRS in Amharic. To mention one, an Amharic dictation machine is one of the application areas.

Dictation has been a common application area of ASRSs for a long period. It includes medical transcriptions, legal and business dictation, as well as general word processing. This is especially helpful for many people who have difficulty in typing due to physical limitations and for those who are very slow in typing. We can even consider using it to allow the illiterates to dictate their ideas.

In some cases, special vocabularies and tuning the language model may be required to increase the coverage of the system. The application of our speech recognizer as a dictation machine may only require a few modifications, in addition to what we have recommended above.

Bibliography

[1] Bahl, R. Frederick Jelinek, and Robert L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 5(2):179-190. In A. Waibel and K.F. Lee, editors, Readings in Speech Recognition, pp 308-319. Morgan Kaufmann, 1990.

[2] Baye Yimam. 1986. "የአማርኛ ሰዋሰው". Addis Ababa ት. መ. ማ. ማ. ድ.

[3] Baye Yimam and TEAM 503 students. 1997. "ፊደል እንደገና" Ethiopian Journal of Languages and Literature 7(1997): 1-32.

[4] Bender, L.M. and Ferguson C. 1976. The Ethiopian Writing System, in Bender, M. et al (eds.), Languages in Ethiopia. London: Oxford University Press.

[5] Beyene, Berhanu; Manfred Kudlek; Olaf Kummer; Jochen Metzinger. 1997. The ethiop package. Fachbereich Informatik, Universität Hamburg. `ftp://ftp.dante.de/tex-archive/languages/ethiopia/ethiop/`

[6] Brugnara, F., Falavigna, D. and Omologo, M. 1993. Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models. Speech Communication, 12, 4, 357-370

[7] Byrne, William et al. 2000. Morpheme Based Language Models for Speech Recognition of Czech In: Proceedings of TSD 2000 pp211-216. Springer-Verlag Berlin.

[8] C'ernocky',Jan. 2002. Temporal processing for feature extraction in speech recognition. `http://www.fit.vutbr.cz/~cernocky/publi/2002/habil.pdf`

[9] Chen, Scott Shaobing and Peter DeSouza. 1997. Speaker Adaptation by Correlation (ABC). In the Proceedings of the DARPA Speech Recognition Workshop. `http://www.nist.gov/speech/publications/darpa97/`

[10] Cowley, Roger, et al. 1976. The Amharic Language-Description. In Language in Ethiopia. Edited by M.L. Bender, J.D. Bowen, R.L. Cooper, and C.A. Ferguson. London: Oxford University Press.

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

[12] Frasen, J., et al. 1994. WSJCAM0 Corpus and Recording Description. Technical Report: CUED/F-INFENG/TR.192. Cambridge University, Engineering Department. Cambridge.

[13] Fukada, T.; T. Yoshimura, and Y. Sagisaka. Automatic generation of multiple pronunciations based on neural networks. Speech Communication 27:63–73, 1999. http://citeseer.ist.psu.edu/fukada99automatic.html

[14] Ganapathiraju, Aravind; Jonathan Hamaker; Mark Ordowski; and George R. Doddington. 1997. Joseph Picone. SYLLABLE-BASED LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION. http://www.cavs.msstate.edu/hse/ies/publications/journals/ieee_sap/2001/syllable_asr/paper_v7.pdf.

[15] Getachew Haile. 1967. The Problems of the Amharic Writing System. A paper presented in advance for the interdisciplinary seminar of the Faculty of Arts and Education. HSIU.

[16] Gibbon, Dafydd Roger Moore and Richard Winski. ed. 1997. Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Walter de Gruyter Publishers.

[17] Hayward, Katrina and Richard J. Hayward. 1999. Amharic. In Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge: The University Press.

[18] Hu, Zhihong; Johan Schalkwyk; Etienne Barnard; and Ronald Cole. 1996. Speech recognition using syllable like units. Proc. Int'l Conf. on Spoken Language Processing (ICSLP), 2:426-429.

[19] Huckvale, M A and A C Fang. 2001. EXPERIMENTS IN APPLYING MORPHOLOGICAL ANALYSIS IN SPEECH RECOGNITION AND THEIR COGNITIVE EXPLANATION. Institute of Acoustics Workshop on Innovation in Speech Processing, Stratford-on-Avon. www.speech.sri.com/cgi-bin/run-distill?papers/icslp2004-arabic-lm.ps.gz www.phon.ucl.ac.uk/home/mark/papers/WISP2001-morph.pdf

[20] Jelinek, F. 1990. Self-organized language modeling for speech recognition. In A. Waibel and K.F. Lee, editors, Readings in Speech Recognition, pp 450-506. Morgan Kaufmann.

[21] Junqua, J.-C. and J.-P. Haton 1996. Robustness in Automatic Speech Recognition: Fundamentals and Applications. Boston: Kluwer Academic Publishers.

[22] Kanokphara, Supphanat; Virongrong Tesprasit and Rachod Thongprasirt. 2003. Pronunciation Variation Speech Recognition Without Dictionary Modification on Sparse Database, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003, Hong Kong).

[23] Kim, Woosung and Myoung-Wan Koo. 1999. A Korean speech corpus for train ticket reservation aid system based on speech recognition. In Proc. of European Conference on Speech Communication and Technology. `http://www.cs.jhu.edu/~woosung/ps/icsp97.ps`.

[24] Kare, Sjölander. 2001. Automatic alignment of phonetic segments. Lund University, Dept. of Linguistics, Centre for Speech Technology, Dept. of Speech, Music, and Hearing, KTH. `http://www.ling.lu.se/disseminations/pdf/49/bidrag36.pdf`.

[25] Kinife. 2002. Sub-word Based Amharic Word Recognition: An Experiment Using Hidden Markov Model (HMM), M.Sc Thesis. Addis Ababa University Faculty of Informatics. Addis Ababa.

[26] Lee, Kai-Fu. 1989. Automatic speech recognition : the development of the SPHINX system. Boston; London: Kluwer Academic.

[27] Lee, C-H., Gauvain, J-L., Pieraccini, R. and Rabiner, L. R. 1992. Large vocabulary speech recognition using subword units. Proc. ICSST-92, Brisbane, Australia, Dec. 1992, pp. 342-353.

[28] Leslau, W. 2000. Introductory Grammar of Amharic, Wiesbaden: Harrassowitz.

[29] Leslau, W. 1995. Reference Grammar of Amharic, Wiesbaden: Harrassowitz.

[30] Martha Yifiru. 2003. Application of Amharic speech recognition system to command and control computer: An experiment with Microsoft Word, M.Sc Thesis. Addis Ababa University Faculty of Informatics. Addis Ababa.

[31] Markowitz, Judith A. 1996. Using speech Recognition. New Jersey: Prentice Hall PTR. Lukaszewica, Konrad and Matti.

[32] Rabiner, L. and Juang, B. 1993. Fundamentals of speech recognition. Englewood Cliffs, NJ.

[33] Rabiner, L. R. May 1999. "Speech recognition in machines"in The MIT Encyclopedia of the Cognitive Sciences, Wilson, R. and Keil, F. K. (Ed.), MIT Press. `http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/mitecs_paper.pdf`

[34] Radova, Vlasta and Petr Vopalka. 1999. Methods of Sentences Selection for Read-Speech Corpus Design. In V. Matousek et al. (Eds.) TSD'99, LNAI 1692, pp. 165-170. Berlin Heidelberg: Springer-Verlag

[35] Rodman, Robert D. 1999. Computer Speech Technology. Boston and London : Artech House.

[36] Roukos, Salim. 1996. Language Representation. In Survey of the State of the Art in Human Language Technology edited by Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen and Victor Zue. `http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html`

[37] Schiel, Florian and Christoph Draxler 2004 The Production of Speech Corpora. `http://www.phonetik.uni-muenchen.de/Forschung/BITS/TP1/Cookbook/Tp1.html`

[38] Schultz T., Rogina I.: Acoustic And Language Modeling of Human and Nonhuman Noises for Human-toHuman Spontaneous Speech Recognition, appeared in: The Proceedings of the ICASSP 1995 `http://citeseer.ist.psu.edu/schultz95acoustic.html`

[39] Solomon Birihanu. 2001. Isolated Amharic Consonant-Vowel (CV) Syllable Recognition, M.Sc Thesis. Addis Ababa University Faculty of Informatics. Addis Ababa.

[40] Svendsen, Torbjrn; NTH and Knut Kvale. 1995. Automatic alignment of phonemic labels in continuous speech. Telenor Research COST 249, Nancy, March 6-7, 1995. `http://www.elis.rug.ac.be/ELISgroups/speech/cost249/report/references/papers/sve95a.pdf`

[41] Tadesse Beyene. 1994. The Ethiopian Writing System. Paper presented at the 12^{th} International Conference of Ethiopian Studies, Michigan State University.

[42] Tang, M. 2005. Large Vocabulary Continuous Speech Recognition Using Linguistic Features and Constraints. MIT Department of Electrical Engineering and Computer Science.

[43] Titov, EG 1976. The Modern Amharic Language. Moscow: Nauka Publishing House.

[44] Ullendorff, Edward. 1965. An Amharic Chrestomathy.London: Oxford University Press.

[45] Vergyri, D.; K. Kirchhoff; K. Duh, and A. Stolcke, "Morphology-based language modeling for Arabic speech recognition," in Proc. Intl. Conf. Spoken Language Processing, (Jeju, Korea), pp. 2245–2248, October 2004. `www.speech.sri.com/cgi-bin/pubs/glimpse_to_html.pl`

[46] Woodland, Phil C. 2001. Speaker adaptation for continuous density HMMs: A Review. Invited Lecture, In Adaptation-2001, 11-19.

[47] Wu, Su-Lin. 1998. Incorporating Information from Syllable-length Time Scales into Automatic Speech Recognition. PhD thesis, University of California, Berkeley, CA.

[48] Young, Steve. 1996. Large Vocabulary Continuous Speech Recognition: A Review. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp 3–28, Snowbird, Utah, December 1995. IEEE. `http://citeseer.ist.psu.edu/article/young96large.html`

[49] Young, Steve; Dan Kershaw; Julian Odell and Dave Ollason. 2002. The HTK Book.

[50] Zegaye Seyifu. 2003. Large vocabulary, speaker independent, continuous Amharic speech recognition, M.Sc Thesis. Addis Ababa University Faculty of Informatics. Addis Ababa.

Appendix A

Documentation of Amharic Corpus

This is the documentation for the Amharic Speech Corpus (ASCo) which is recorded in April - June 2003, and extended in February 2004 as a part of this Ph.D work.

ASCo contains the recording sessions of 124 Amharic native speakers. The training set is read by 100 speakers. Out of them seventy speakers read a list of 100 prompts, ten speakers read a list of 145 prompts, and twenty speakers read 120 prompts for the training set. Twenty-four speakers read a list of 139 prompts for the test sets (development and evaluation) and speaker adaptation set.

All of the speakers read the prompts from a screen. All of the prompt sentences are included in the corpus with the encoding that is given in Appendix C, page 111 and in the ethiop fonts.

The speakers are recruited on the basis of their age, sex and reading ability. (The age and sex distribution of the readers is given in Table A.1, page 107.) For the extension of the corpus speakers are recruited on the basis of their Amharic dialect.

The recording is made with a high quality close speaking headset microphone that has a noise canceling capability. During recording, the microphone is placed at about 5cm on the left-side of the speakers' mouth. The researcher is present in all the recording sessions to control the recording set-ups.

Age Range	Training set		Test sets	
	Male	Female	Male	Female
Speakers of the Addis Ababa dialect				
18-23	18	18	3	3
24-28	12	12	3	3
29-40	5	5	3	3
Older than 40	5	5	1	1
Total	40	40	10	10
Speakers of the other four dialects				
18-23	10	3	4	
24-28	6	1		
Total	16	4	4	
Grand Total	56	44	14	10

Table A.1: Age and Sex Distribution of the Readers.

The office in which the speech is recorded is not soundproof and is located in the Addis Ababa University, building of the Institute of Language Studies (ILS).

The speech is recorded using a Toshiba laptop that has a Pentium 4 processor and an Intel 82801DBAC'97 sound card. All of the internal sounds on the laptop are recorded with the speech. This set of noise makes up the constant noise of the corpus. There are also some external variable noise sources, like doors of the nearby offices, and people outside the recording office who have rarely created very loud sounds that go beyond the canceling capacity of the microphone.

The speech is transcribed at word level and annotated at word and syllable level semi-automatically. There are label files at word, syllable and phone level. The label files, which are in the HTK format, are included in the corpus.

The utterance identification is used as a file name of the speech files and label files with extension wav and lab, respectively. Each file name is composed of four parts:

1. The data group - the first two alpha-numerals,
2. The reader code - the next three (two for test and adaptation sets) digits,

3. The prompt/label code - the following three digits, and
4. The extension - the remaining three letters after the dot.

For example:

1. In the file tr001002.wav:
   (a) tr is the data group and stands for training,
   (b) 001 is the reader code and represents the first reader in the training group,
   (c) 002 is the prompt/label code and represents the second prompt for this reader,
   (d) wav is the extension that shows the file is a speech file in its wave format.
2. In the file d501021.wav:
   (a) d5 is the data group and stands for development test set with 5,000 words,
   (b) 01 is the reader code and represents the first reader in the test group,
   (c) 021 is the prompt/label code and represents the twenty-first prompt for this reader,
   (d) wav is the extension that shows the file is a speech file in its wave format.

The data is structured in the following way: The files of the training speech are in the directory `training` while the files of the test and adaptation speech are saved in the directory `test`. Within these directories, there are directories for each speakers named after the speaker code and the speech files of the corresponding speaker are saved in these sub-directories.

The label files and any other documentation files, including this one, are saved in a directory called DOC. All of the files of different programs that are written for the purpose of text processing and any other automatic pre- and post-processing works are saved in the directory called Scripts.

Appendix B

Configuration of Speech Feature Extraction

```
SOURCEKIND=WAV
SOURCEFORMAT=WAV
SOURCERATE=625
TARGETFORMAT=HTK
TARGETKIND=MFCC_0_D_A
TARGETRATE=100000
SAVECOMPRESSED=TRUE
SAVEWITHCRC=TRUE
WINDOWSIZE=250000.0
USEHAMMING=TRUE
PREEMCOEF=0.97
NUMCHANS=26
CEPLIFTER=22
NUMCEPS=12
ENORMALISE=TRUE
```

The first four elements of the configuration specify that source files of the speech are in wave format of sixteen sample rate and to be converted to the HTK format. The others specify that the target parameters are to be MFCC, using C_0 as the energy component, the frame period is 10 ms (HTK uses units of 100ns), the output should be saved in compressed format, and a crc checksum should be added. The FFT should use a Hamming window of size 25 ms (which is the frame size), and the signal should have first order pre-emphasis applied, using a co-efficient of 0.97. The filter-bank should have twenty-six channels and 12 MFCC co-efficients should be the output. The variable ENORMALISE is by default true and performs energy normalization on recorded audio files (Young 2002).

Appendix C

Encoding of Amharic CV Syllables

		a	u	i	A	E	e	o	ua	ui	uA	uE	ue
	IPA	ə	u	i	a	e	ɨ	o	^{w}ə	^{w}i	^{w}a	^{w}e	^{w}ɨ
h	h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ	ኈ	ኊ	ኋ	ኌ	ኍ
l	l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ			ሏ		
m	m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ			ሟ		
r	r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ			ሯ		
s	s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ			ሷ		
S	ʃ	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ			ሿ		
q	q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቈ	ቊ	ቋ	ቌ	ቍ
b	b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ			ቧ		
v	v	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ			ቯ		
t	t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ			ቷ		
c	tʃ	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ			ቿ		
n	n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ			ኗ		
N	ɲ	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ			ኟ		
H	?	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ	ኧ				
k	k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኰ	ኲ	ኳ	ኴ	ኵ
K	h	ኸ											
w	w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ					
z	z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ			ዟ		
Z	ʒ	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ			ዧ		
y	j	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ					
d	d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ			ዷ		

Cont.

		a	u	i	A	E	e	o	ua	ui	uA	uE	ue
	IPA	ə	u	i	a	e	ɨ	o	^{w}ə	^{w}i	^{w}a	^{w}e	^{w}ɨ
D	dʒ	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ			ጇ		
g	g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ	ጐ	ጒ	ጓ	ጔ	ጕ
T	t‘	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ			ጧ		
C	tʃ‘	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ			ጯ		
P	p‘	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ			ጿ		
x	s‘	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ			ጿ		
f	f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ			ፏ		
p	p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ					

Index

Ich erkläre hiermit an Eides statt, dass ich diese Arbeit selbst verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe.

Hamburg, im Dezember 2005

Solomon Teferra Abate