ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
DEPARTMENT OF INFORMATION SCIENCE

Speaker Independent Continuous Amharic Speech Recognizer: An Experiment Using Hybrid (HMM-ANN) System

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

BY

Hussien Seid

JULY 2004

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION STUDIES FOR AFRICA

Speaker Independent Continuous Amharic Speech Recognizer: An Experiment Using Hybrid (HMM-ANN) System

BY

Hussien Seid

Signature of the Board of Examiners for Approval

_____	_____
_____	_____
_____	_____
_____	_____

DEDICATION

This thesis is dedicated to all people who have had their own share of contribution throughout my life.

ACKNOWLEDGEMENTS

My first and most thanks go to my Lord Allah, the Gracious and most Merciful. Without His help, not only this but also any other part of my life would not have been successful.

Next I would like to thank my advisor Dr. Björn Gambäck, Ato Kinfe Taddesse and Ato Zegaye Seifu for their constructive comments and encouragement throughout this work. Another special thanks for Marek F. and Clemente Fragoso Eduardo for their real and dedicated help on fixing the CSLU Toolkit implementation problems on my machine. I thank you all, for your immediate response in helping me out whenever I got in trouble.

This research couldn't have come into being without the help of Ato Solomon Tefera. I thank you for the corpus that you have given me. Thank you.

I am also grateful to my family, specially my mother Fatuma Ahmed and my wife Husniya Damtew, who have rendered me all their care, love and encouragement. I would like to mention my uncle Ato Muhamed Ahmed. I am every thing today because yesterday he was at my back.

Finally, I would like to extend my heart-felt gratitude to all people who have had their own share of contribution throughout my life. My classmates, brothers and friends who have been sick for my cases on this work, I have no word to express what you were for me. Simply I love you all.

Table of Content

List of Figures and Tables

List of Annexes

ACRONYMS

ASR – Automatic Speech Recognition

HMM – Hidden Markov Model

CSLU – Center of Spoken Language

ANN – Artificial Neural Network

MLP – Multilayer Perceptron

MFCC – Mel-Frequency Cepstral Coefficient

Abstract

Automatic speech recognition is a part of natural language processing which deals with making computers hear human speech and to take action according to what they heard. In line with this concept, Speaker Independent Continuous Amharic Speech Recognition is investigated to check the results that can be reached by an ANN/HMM hybrid approach.

To implement the Amharic ASR system, 100 speakers, each speaking ten different sentences have been modeled with the help of the CSLU Toolkit. The model is constructed at a context dependent phoneme sub-word level. A promising result of ***78.56%*** *word accuracy and* ***44.07%*** *sentence recognition rate has been achieved for speaker dependant test and* ***74.28%*** *word accuracy and* ***43.70%*** *sentence recognition rate for speaker independent test. These are the best figures reported so far for speech recognition for Amharic.*

Chapter I General Overview

1. Background

No time has been observed that man stop striving to change and make various social and technological developments. Particularly in the field of information technology, to make machines act like humans is incredibly bearing fruit. One aspect of such attempts is letting machines recognize human speech in such a way that they will be able to understand and respond to a human friend.

Developing such machines is very important because speech interfacing in the user's own language is an ideal means of communication for being the most natural, flexible, efficient and convenient option that frees hands and eyes for other tasks. Moreover, speech is the ultimate, ubiquitous method of communication and it is how we should be able to interact with computers.

One research area towards developing intelligent machines that can naturally communicate with human beings is automatic speech recognition. Automatic Speech Recognition (ASR) is the way by which machines are made to have the ability to take human speech as an input and produce a transcription of that speech as an output. More specifically, as described by Junqua and Haton (1996), it is the process of "decoding of the information conveyed by a speech signal and its

transcription into characters." In this research, the resulting characters of speech recognition system are used to represent the Amharic character.

The first significant human effort towards the development of functional ASR system was undertaken in the 1970s (O'Shaughnessy, 2000). A lot has been done for various languages like English, Dutch, Chinese and so forth. Richard Lippman (1997) says that ASR currently performs about an order-of-magnitude worse than the human listener. On the other hand, in our country only small but remarkable investigations have been observed for the past few years. That is why no operational ASR system for the languages of our country is available in the market place.

2. *Automatic Speech Recognition (ASR)*

Practically there are many different forms of ASR ranging from direct voice input (DVI) devices which are primarily aimed at voice command and control; to large vocabulary continuous speech recognition (LVCSR) systems which are used for form filling or voice–based document creation (Moore, 2002). Rabiner and Schafer (1987) formulated eight specification attributes to classify or categorize ASR:

1. Types of speech—discrete words, continuous speech, etc
2. Number of speakers—single, multiple, unlimited, etc

3. Types of speakers—cooperative, casual, male, female, child etc
4. Speaker environment—sound proof booth, computer room, public place
5. Transmission system—high quality microphone, close talking microphone, telephone
6. Type of system training—no training, fixed training set, continuous training
7. Vocabulary size—small (1-1,000 words), medium (1,000-10,000 words), large (more than 10,000) vocabulary
8. Speech input format—constrained text, free spoken format

In any way, based on the flow of input speech, size of vocabulary and the type of speaker modeling (Solomon, 2001), ASR systems can be classified generally based on the following parameters: (Kinfe, 2002)

- Discrete word versus Continuous speech
- Speaker Dependent versus Speaker Independent
- Context Sensitive versus Context Insensitive
- Large Vocabulary versus Small Vocabulary

2.1. Discrete Word versus Continuous Speech

Discrete word recognition systems recognize only separate words. The speaker should pause briefly between the words to be recognized (Zue and Cole, 1995). As indicated by Lea (1982) this artificial pause

helps to identify word boundaries easily. In addition, the co-articulation effects on the pronunciation of each word are minimized. Markowitz (1996) also indicated the fact that the pauses allow the processor time to accomplish its analysis. Such types of recognizers are the easiest to develop and can perform efficiently for command and control application but introduce lower speech rate and some user inconveniencies.

To the contrary, in continuous speech recognizers, a user speaks naturally without artificial pauses between words. Hence continuous speech recognition is more difficult than discrete word recognition. This is because of the following three properties of continuous speech.

- Word boundaries are unclear in continuous speech;
- Coarticulation effects are much stronger in continuous speech;
- Content words are often emphasized while function words are poorly articulated (Lee, 1989).

2.2. Speaker Dependent versus Speaker Independent

Speaker dependent systems recognize speech from those people whose speech is used during the development of the recognizer. Since recognition is done on the people used for training the model, it avoids most of the speaker variability problem (Junqua and Haton, 1996).

Speaker-independent systems, on the other hand, are capable of recognizing speech from any new speaker. However, designing such

system is difficult since most parametric representation of speech are highly speaker dependent, and a set of reference patterns suitable for one speaker may perform poorly for another speaker (Lee, 1989). Speaker dependent systems achieve better recognition performances than speaker-independent systems.

Either speaker dependent or speaker independent can be integrated with discrete word or continuous speech recognition systems. Based on this integration, ASR systems can be categorized as:

- Speaker dependent isolated word ASR system
- Speaker dependent continuous speech ASR system
- Speaker independent isolated word ASR system
- Speaker independent continuous speech ASR system

From these four categories, the Speaker independent continuous speech ASR systems are the ultimate goal of speech recognition research (Kinfe, 2002).

2.3. Context sensitive versus context insensitive ASR systems

In most speech recognition systems, in order to increase the accuracy, a language model or artificial grammar that approximates the natural language will be applied to restrict the permissible combination of words. Such systems are called context sensitive speech recognition system. However, on the other hand, if no check up system for

contextual restrictions is incorporated then the system is said to be a context insensitive ASR system.

2.4. Small vocabulary versus large vocabulary ASR systems

This classification is based on the size of vocabulary used by the system. Lee (1989) and Rabiner et al (1987) indicated that large vocabulary means a vocabulary of about 100 words or more. Nowadays a large vocabulary is considered to have about 10,000 words, at least. As the vocabulary size is increased, a number of problems arise such as inherent confusability of words, searching difficulties in large vocabulary systems, etc

It is also difficult to model each and every word in large vocabulary systems, because neither the training nor the storage is available. Thus, to develop a large vocabulary system we have to find good sub-word units although sub-word units usually lead to degraded performance since they can not capture co-articulatory (inter-unit) effects (Martha, 2003). Lee (1989) also said that good units of speech should be trainable, well defined and relatively insensitive to context. On the other hand, small vocabulary recognition systems are easier to design and are successful as compared to large vocabulary systems.

3. *Approaches to ASR systems*

In general it can be said that there are four basic approaches (though various writers may use different categorization) for developing speech recognition systems – acoustic-phonetic, template matching (pattern recognition), stochastic approaches, and Artificial Intelligence (AI) (Zegayie, 2003).

3.1. Acoustic-phonetic

Acoustic-phonetic is the oldest, most straight-forward and thoroughly researched method (Reidener, 1999). The approach assumes that the rules governing the phoneme variability are relatively simple and easily learnable. This approach has limitations like absence of well-defined automatic procedures and standard linguistic ways of labeling the training speech. Due to these, the acoustic – phonetic speech recognition is no longer considered as the most interesting approach, but its underlying ideas are still used in the artificial intelligence based recognizers (Ibid).

3.2. Template Matching

Template matching is one form of pattern recognition that represents speech data as sets of feature vectors called templates. Each word or phrase in an application is stored as a separate template. Spoken input by end users is organized into templates prior to performing the recognition process. The input is then compared with

stored templates. Template matching is performed at the word level and contains no reference to the phonemes with in the word.

Template matching approaches use a dynamic time warping (DTW) method (Bush and Copec, 1985) and can perform best for discrete word speech recognition. Of course it is also possible to use for continuous or connected speech recognition with a very large size vocabulary, but this is definitely inefficient.

3.3. Stochastic

The stochastic approach entails the use of probabilistic models to deal with uncertain and incomplete information found in speech recognition. In this approach the most and well establish stochastic model, the Hidden Marcov Model (HMM) is used. During the last few decades HMM has been the most recognized model especially for large vocabulary speaker independent continuous ASR systems. (Rigoll et al, 1996). It is a stochastic generative process that is particularly suited to model time–varying patterns such as continuous speech (Moore, 2002). Still, there are a number of tasks that can cause considerable problems for current HMM techniques (Tuerk, A., 2001).

3.4. Artificial Intelligence

The main idea of Artificial Intelligence (AI), the fourth approach, is to collect and employ knowledge from a number of sources in order to

solve the problem in question. The knowledge sources are wide ranging from the fields of acoustic, lexical, syntactic, semantic and pragmatic knowledge (Rabner and Juang, 1993).

The most important techniques in this approach are the use of an expert system for segmentation and labeling of the acoustic signal, learning and adaptation over time, and the use of artificial neural networks (ANN) for the distinction between similar sound classes and learning the relations between all known inputs and phonetic data. The use of ANN does not outperform the use of HMM on benchmark tests (Borland et al, 1996).

The neural networks can represent a separate structural approach to speech recognition or can be regarded as an implementation architecture possibly incorporated in any of the three classical speech recognition approaches. Based on this concept the hybrid of HMM and ANN has been developed to hit the unique advantages of both ANN and HMM (Bilmes et al, 1999).

In the hybrid approach significant improvements over purely HMM based systems have been obtained by using ANN in some specific subsystems of the speech recognizer (Beaufays et al, 2001). It is presented in the last few years as a better alternative to HMM based systems for large vocabulary speaker independent continuous ASR

systems (Almeid et al, 1996) and the most successful approach (Rottland et al, 1996). And this is the method to be used in this research too.

4. Benefits and Applications of Speech Recognition

The naturalness of the speech makes many researchers to be committed in developing automatic speech recognition system (Kinfe, 2002). Any task that involves interfacing with a computer can potentially use Automatic Speech Recognition. To sample out some but not all, the following applications of ASR can be mentioned (Zegaye, 2003).

- Telephone systems (automatic phone dialing)
- Audio typewriter (dictation systems) like automatic typewriter that is activated and work by voice (speech)
- As an assistive technology, handicapped people who cannot use keyboard and other pointing devices can be made to use computers, telephones, fax machines, etc (Rodman, 1999)
- In education for Computer Aided Instruction systems
- Data entry and retrieval by voice from remote and busy work stations
- For seminar scribers, News agencies and meetings—the speech of a speaker can be directly transcribed onto paper, while he/she is speaking.
- Medical journals–according to Markowith (1996), "using speech recognition to generate reports – actually takes less of the doctor's time, requires no transcription, and is available immediately. So that when the patients get wheeled into the operating room the report is there rather than two weeks latter."

- Courts and legal institutions do benefit from the application of ASR system. Direct and proper transcription of the speech from the defendant, the testimonials and letter of the judges – on a case, for example – is indispensable and is possible by an ASR.
- In industrial and military and health care environments to control and command different machines
- In some multimodal interfacing applications, integrating speech and pointing devices are more efficient than graphical interface without speech

5. *Statement of the Problem and Justification*

As indicated earlier, the ultimate aim of ASR research is the development of an ASR system that recognizes anything anyone says on any topic in any situation. To achieve this goal, research has been conducted for about 50 years and various systems have been developed for different languages, although the ultimate goal is not achieved till now due to the various challenges associated with speech recognition. Researchers are still struggling with the various challenges so as to achieve the goal.

However, in our country, research in this area started only recently. As a result there are few researches on speech recognition in particular and speech technology in general. Specifically, research conducted on speech technology for Amharic language has been very limited. Laine (1998) made a valuable effort to develop an Amharic text-to-speech synthesis system. Solomon (2001) and Kinfe (2002) developed

Amharic consonant vowel (CV) syllable and isolated word recognizers, respectively. With their research, both made efforts to develop speaker dependent and speaker independent systems. Zegaye (2003) and Martha (2003) have continued by working on a large vocabulary speaker independent continuous ASR system and ASR for command and control applications, respectively. However, there is still a lot of work to be done towards achieving a full-fledged automatic Amharic speech recognition system.

As the fact that Amharic is a dominant language in Ethiopia and the official language of the Federal Government of Ethiopia (Worku, 1997) and most commonly learned second language throughout the country, it is very important to make further contributions in this area. Therefore, this study aims at investigating and testing out the possibility of developing large vocabulary, speaker independent continuous Amharic speech recognition systems using a hybrid of HMM and ANN systems. This will consummate the endeavor towards the construction of continuous Amharic speech recognition and the result will be an input to the development of real applications.

6. *Objectives*

Having the back ground this much, in this research the following general and specific objectives are aimed to attain

6.1. General Objective

The general objective of the study is to examine and demonstrate the performance of a hybrid HMM/ANN system for a speaker-independent continuous Amharic speech recognizer.

6.2. Specific Objectives

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

- ✓ Conducting a literature review on Automatic Speech Recognition in general and on how to develop a recognizer for the Amharic language.
- ✓ Read and learn the CSLU toolkit which is used in this research.
- ✓ Identify speech units suitable for the Amharic speaker-independent continuous speech recognizer.
- ✓ Collect Speech data appropriate to construct Amharic ASR System
- ✓ Mark up the collected speech data at the identified speech unit level
- ✓ Build the recognizer.
- ✓ Test the performance of the developed recognizer.
- ✓ Forward conclusion and recommendation for further study.

7. *Methodology*

The following methods have been employed in conducting the proposed study.

7.1. Literature Review

Exhaustive literature review has been carried out to investigate the underlying principles/theories of the various approaches, techniques and tools that were employed in the research. In addition, literature on the Amharic language, especially those dealing with the phonetic/triphone features, were reviewed. Moreover, to learn what others have done in the area and to better understand the problem, a comprehensive investigation of available empirical literature on automatic speech recognition has been carried out.

7.2. Data Collection and Preparation

It is obvious that the data required for this research is Amharic speech taken from people of different age, sex and if possible linguistic background. Then this speech data has been processed using the available laboratory equipment of the faculty, Dell Pentium IV processor with 2.0GHz Proccessor speed and 256Mb RAM. The speech data has then been marked up and divided into three parts: training, development and test data sets. The training data set has been used to train the recognizer while the development and test data set have been used to evaluate and test the performance of the recognizer.

7.3. Modeling Technique

In this study, CSLU Toolkit is used as a main Tool kit to preprocess the speech data and to construct, evaluate and test the

Amharic Automatic Speech Recognition System, which is developed just guided by the manual given with the Toolkit. To construct the vocabulary and other text files needed for development of the system and to view some output files of the commands of the toolkit, some text editor soft wares like WordPad and WinWord have been used. The command prompt of the Microsoft Window, ***cmd.exe*** was used to call commands of CSLU Toolkit.

7.4. Testing Technique

The recognition system must be evaluated or tested for its accuracy. To test the recognition system the test data set has been used and then the accuracy has been reported in terms of the three common recognition errors: deletion, insertion and substitution. Tables and percentages will be used to present the accuracy report.

8. *Application of Results*

Any Ethiopian who can speak Amharic can use the result of this research. It is specifically important for groups interested in developing Amharic software like Dictation Systems.

Chapter II Linguistic/Phonetic Background

1. Introduction

ASR systems are very language dependent. This means that one ASR system works well only for the language it is developed, because it integrates phonetic information of the language. For example, an ASR system needs to have pronunciation for every word it can recognize (Jurafsky, 2000) and produce the corresponding symbolic representation. This shows that at least some linguistic review is very important, to build an ASR system. The main requirement here is a comprehensive analysis of speech as a system which is a stream of acoustic signals. Moreover "Understanding of how speech sounds are produced by the human vocal apparatus and how speech is perceived by human auditory system are also vital" (Kinfe, 2002).

Natural languages often have two parts: spoken and written language. In spoken language, the spoken words are composed of small units of speech, called phones (Jurafsky et al, 2000) or phonemes (The Random House Dictionary for English Language, 1973) whereas in written language, the smallest unit is the character. Phonemes are sets of sounds that are the basic building blocks of any language (Pablo Zegers, 1998). In some languages, which have logographic writing systems a whole word can be represented by a single character. There

are other languages that follow syllabic or phonemic writing systems, in which symbols (characters) are to represent phones or sounds that make up the word (ibid).

2. *Taxonomy of Phonemes*

Phonemes are produced by expelling airflow from the lungs through different arrangement of what we call voice box, and vocal tracts, which consists of oral and nasal tracts. Based on the way phonemes are produced, they are divided into two classes: consonants and vowels. Consonants are produced with substantial obstruction of the oral cavity, whereas vowels are produced with a relatively free air flow (Clark et al, 1985).

Based on the place of articulation, manner of articulation and the manner of voicing, consonants can be further classified as follows (Jurafsky et al, 2000) plus the figure given in Figure 1. shows the place of articulations clearly.

Place of Articulation

Labial: consonants whose main restriction is formed by the two lips coming together have a bilabial place of articulation.

Dental: sounds that are made by placing the tongue against the tooth are dentals

Alveolar: the alveolar ridge is the portion of the mouth just behind the upper teeth

Palatal: the roof of the mouth (the palate) rises sharply from the back of the alveolar ridge to produce the so called Platal consonant.

Velar: the velar or soft palatal is a movable muscular flap at the rear of the roof of the mouth. So consonants produced at this point are known as Velar

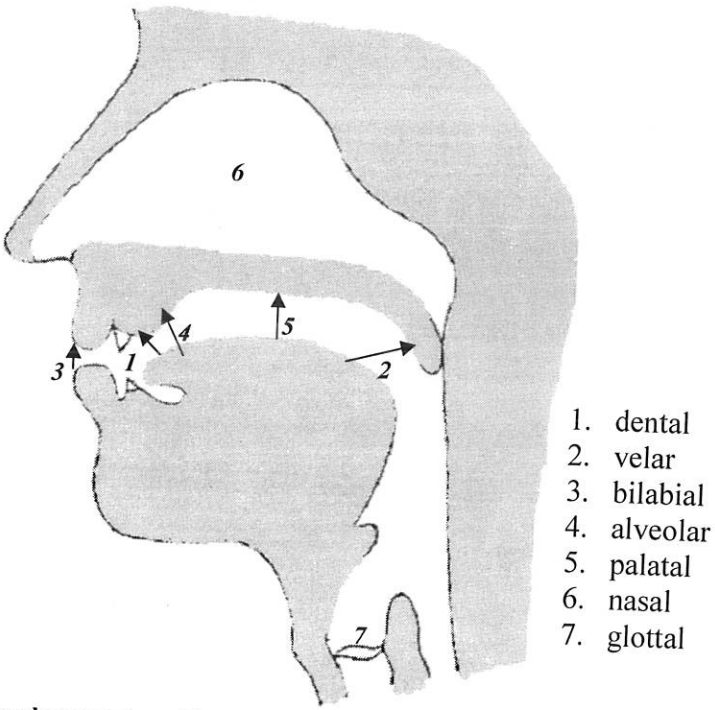


Figure 1: Vocal apparatus of human beings showing place of articulations

Manner of Articulation

Stop: A (glottal) stop is a consonant in which the airflow is completely blocked for a short time and often followed by an explosive sound as air is released

Nasal: the nasal sounds are made by lowering the velum and allowing air to pass into the nasal cavity.

Fricative: in this case, airflow is constricted but not cut off completely.

Approximant: In approximants, two articulators are close together, but not close enough to cause turbulent airflow.

Tap: A tap or flap is a quick motion of the tongue against the alveolar ridge.

Manner of Voice

Voiced: sounds made with the vibrating of the vocal cord.

Voiceless: sounds made without the vibrating of the vocal cord

Similarly vowels also can be classified by the position of the articulators as they are made. The most relevant parameters for vowels are what are called vowel height, position and the shape of the lips

(Solomon, 2001). Hence, a vowel can be high, mid or low based on the height of the tongue body with respect to base of the mouth. Based on the position of the tongue a vowel can be produced at the back, front or center of the mouth.

3. The Amharic Writing System

The Amharic writing system is based on its own character set called 'fidEl' which consists of 275 characters excluding twenty numerals and eight punctuation marks. From the total character set, 238 characters represent consonant—vowel phoneme pairs (CV representation). There are 34 consonants and 7 vowels to make the pairs in a regular format for each consonant called order. The order of the consonants መ /m/ and ሰ /s/, for example, are illustrated below in figure 2.

Order	1st	2nd	3rd	4th	5th	6th	7th	1st	2nd	3rd	4th	5th	6th	7th
'Fidel'	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
Transcription	mE	mu	mi	ma	me	mI	mo	sE	su	si	sa	se	sI	so
CV Representation	ምኧ	ምኡ	ምኢ	ምኣ	ምኤ	ምእ	ምኦ	ስኧ	ስኡ	ስኢ	ስኣ	ስኤ	ስእ	ስኦ

Figure 2. Construction of Amharic character phonemes

There are also some rarely used characters, which are constructed from a consonant and more than one vowel phoneme. They are 37 in number and are not governed by the format mentioned above. Figure 3

and 4 given below also summarize the Amharic consonants and vowels classified according to their phonological characterization respectively.

Figure 3. Consonant Characters of Amharic

Manner of Articulation	Voicing	Place of articulations					
		labial	dental	alveolar	palatal	velar	glottal
Stop	voiced	ቡ/b/		ድ/d/		ግ/g/	
	voiceless	ፐ/p/	ት/t'/			ክ/k/	እ/I/
Fricative	voiced		ቨ/v/	ዝ/z/	ዥ/ž/ ጅ/j/		
	voiceless		ፍ/f/	ስ/s/	ሽ/š/ ች/č/		ሀ/h/
Nasal	voiced	ም/m/					
	voiceless			ን/n/	ኝ/ñ/		
Approximant	voiced			ል/l/	ር/r/		
	voiceless	ው/w/			ይ/y/		
Flap	voiced	ጵ/p'/		ጽ/s'/		ቅ/k'/	
	voiceless			ጥ/t'/	ጭ/c'/		

	front		center		back	
	rounded	Un rounded	rounded	Un rounded	rounded	Un rounded
High		ኢ/i/		እ/I/	ኡ/u/	
Mid		ኤ/e/		ኧ/E/	ኦ/o/	
Low				አ/a/		

Figure 4 Vowel Characters of Amharic

As can be seen from the examples, once the consonants and vowels are recognized, it is very straightforward to pronounce the seven order of consonant. This consequently makes reading an Amharic word easy without even knowing the language or hearing the pronunciation of the word before (assuming you know the character set). It is also easy to write anything in the language without prior knowledge of the language or seeing the spelling before as long as each orthographic symbol is recognized and the word to be written is pronounced correctly. For example, to write the word 'yEmayItamEnI' /Incredible, Unbelievable/ we use የማይታመን.

Relatively to English or other alphabetic languages, Amharic uses a small number of characters to represent a word. Probably this characteristics of the language made some peoples to think the language is phonetic or syllabic (Bender, 1976; Cowely, 1967; Baye, 1986). However, when we analyze each character into its constituent phonemes it will be clear that the language is not syllabic (Tadesse, 1994; Baye, 1997). The word ቋንቋ /language/, for example, has more phonemes than the representing characters.

ቋንቋ /kuanIkua/ = ቀ+ኡ+ኣ+ን+ቀ+ኡ+ኣ

recognizer. Moreover, the same word spoken with different pitches may be recognized as different words and may degrade performance.

Therefore, as a second approach, sub word units are used such as phones and syllables especially for large vocabulary continuous speaker independent ASR systems. The most commonly used units are the phones, i.e. the acoustic realization of phonemes (Beaufays et al, 2001), because they are relatively consistent and trainable and limited in number. As explained earlier, for example, the Amharic language has only 39 phones (Kinfe, 2002).

The third approach is a refined form of the second approach. In this approach, context dependent phone units are used. Usually they are called triphones, which are the acoustic realization of phonemes in the specific context defined by the previous and following phonemes. So for this work also these two approaches are supposed to be used.

CHAPTER III The ANN/HMM Hybrid Model

1. Introduction

As discussed in the first chapter, there are about four approaches to speech recognition systems. Out of these, using the HHM has been observed for impressive recognition performance; however, it has some drawbacks, especially for large vocabulary speaker independent continuous ASR systems.

The structures and algorithms used in HMM provides a rich and flexible mathematical frameworks for building recognition systems. The power full learning and decoding method for temporal sequences with out demanding hand label segmentation is also observed in the structure and algorithms of HMM. In this paradigm different level of constraints like phonological and syntactical constraints are easily accommodated as long as they are expressed in terms of the same statistical formalism. On the other hand, HMM approach made some unrealistic assumptions. The features extracted with in a phonetic segment are, for example, are assumed to be uncorrelated with one another. The training algorithms of HMM are also based on likelihood maximization which leads to poor discrimination. (Bourlard, et al 1998) In general, the pros and cons of this paradigm can be summarized as shown in table 1.

4. Problems in the Amharic Writing System

Though it is so easy to transcribe a word in the writing system, it has some inefficiencies (Bender et al, 1976). In the first place, some CV-phonemes can be represented by more than one character. This redundancy has increased the total character set by ten. The redundant 'fidEl's are shown below; the first character in each group is used in this work.

/sE/ → ሰ, ሠ /hE/ → ሀ, ሃ, ሐ, ሓ, ኀ, ኃ

/s'E/ → ጸ, ፀ /lE/ → ለ, ሉ, ዐ, ዓ

Figure 5. Redundant Characters set Amharic writing system

Another limitation of the Amharic writing system is the lack of a mechanism to mark gemination of consonants. The words /wana/ and /wanna/ are both written as ዋና but give two completely different meanings by geminating the consonant ን /n/.

ዋና /wana/ meaning swimming ዋና /wanna/ meaning main, core

This may require different reference models in the database for the multiple forms of the sound depending on the gemination.

The third problem with the system is an ambiguity with the 6th order characters: whether they are vowelless or not. However, this is not a problem for this work.

5. *Basic Units of Speech*

As defined before, an ASR takes continuous utterances of sound waves to produce some distinct symbols to represent the features of sound waves. The specific sound patterns that can be represented efficiently by symbols are used as units of the speech. In determining the units of speech for an ASR there are three approaches (Kinfe, 2002).

The first approach is using words as a unit of speech because they are the most natural units (ibid). According to Lee (1990) word models are able to capture intra-word contextual effects. Because of this concept, word models are used for small vocabulary systems and provide the best performance.

On the other hand, using word models for large vocabulary continues ASR system has several problems. In the first place, not all words of a language can be modeled, because words may be changed, derived from other languages, invented, inflected, transformed or dissolved out. Secondly, even if it had been possible to count them, it is impractical to train all of them several times consistently for the

Table 1. Pros and Cons of pure HMM

Pros	Cons
• *Provides rich mathematical framework*	• *Provides poor discrimination capacity*
• *Demonstrates powerful learning and decoding methods*	• *Practical requirements for distributional assumptions (e. g. uncorrelated features within an acoustic vector)*
• *Has good abstraction features for sequences, temporal aspects*	• *Typically ignore correlation between acoustic vectors*
• *Has flexible topology for statistical phonology and syntax*	

In recent years an ANN has been used to augment the ASR system. By doing so a new paradigm has been born: the ANN/HHM hybrid model. In this model, the HMM is used as a main structure of the system, and the ANN is used in a specific subsystem of the recognizer. This is the only recent approach with significant performance improvement over the pure HMM model. Because, "*Neural networks show superior pattern classification performance in static classification tasks due to their discriminant learning algorithm, while the HMM structure is able to cope with the temporal alignment properties of the Virterbi algorithm. Therefore a hybrid ANN/HMM system is proposed to benefit from both advantages*" (Reichl et al, 1996).

2. *The Core Concept*

For any ASR system it is assumed that the waveform of a speech signal that comes out of a speaker's vocal apparatus is a representation of the symbols to express the concept or the idea in his mind. Now, for a computer based speech recognition system to recognize the sequence of symbol, W, in the utterance of the speech wave, first the speech waveform should be digitized and parametric features must be extracted from the digitized format. Then the features extracted out of the continuous speech waveform should be converted to a sequence of equally spaced discrete parameter vectors (Young et al, 2002).

Here, the input speech is taken as a temporal sequence of vectors or frames, $O = \{o_1, o_2, o_3, \ldots o_t\}$, which are normally computed at regular time interval say 10ms (Jurafsky et al, 2000).

The frames are used to make the acoustic model which represents the acoustic realization of the speech units based on the underlying assumption that each unit has constant spectral properties (Beaufays, et al, 2001). Then HMM is the most common tool to model the speech unit because it assumes that the sequence of feature vectors is a piecewise stationary process.

The next requirement for a computer-based speech recognizer is to make estimations of the possible symbol sequences that would result in

the information carried by the sequence of vectors. Acoustic models can help such speculations about the probability of the observed symbol sequences given the models. Then a language model can be used to formulate and tell the prior probability of the estimated word sequences (Zegaye, 2003).

In effect, HMM adopt a hierarchical scheme that a sentence is modeled as a sequence of words, W and each word is modeled as a sequence of sub-word units, w_1, w_2, w_3, . . . w_i. (ibid) Moreover it can be defined as a stochastic finite state automation, usually with a left-to-right topology when used for speech (Bourlard et al, 1997) as shown below in Figure 6.

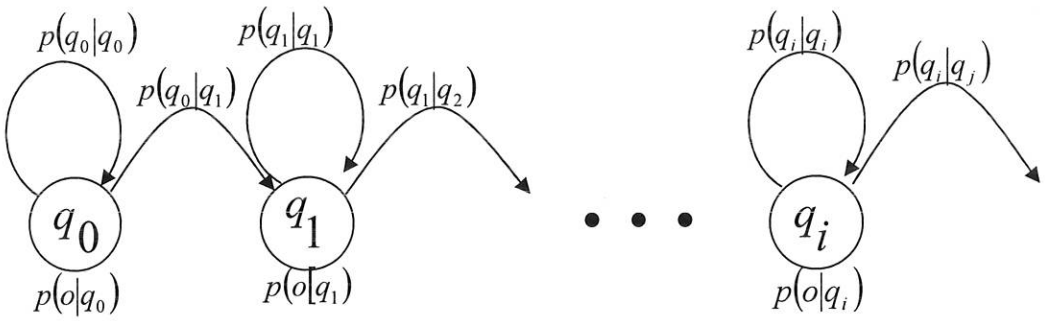


Figure 6 Illustration of HMM states

As it can seen from the figure 6, there are two probabilities to compute for a pronunciation of words: the transition probability $p(q_i|q_j)$ and observation probability $p(o|q_i)$. The transition probability is determined from the pronunciation dictionary of the recognizer. But the

determination of observation probability demands computationally intensive approach (Beaufays et al, 2001). In HMM approach, this probability is determined by using Gaussian probability density function based on the maximum likelihood (ML) technique, which is argued for poor discrimination due to the fact that it maximizes the likelihood of individual acoustic models independently of the likelihood of the other. On the other hand, using multilayer perceptron (MLP) of the artificial neural network (ANN) have scored better discrimination power over the ML; in Beaufays et al (2001) it is expressed as follows:

> *"Neural network classifiers based on MLP provide the simplest architecture for discriminative training: feature vectors extracted from the data frames can be input to an MLP and classified into N classes corresponding to the speech units to be modeled, phones or triphones. When appropriately trained, such an MLP classifier can be shown to minimize the classification error-rate over the training data. This form of training explicitly maximizes the discrimination between the correct output class and its components.*

Hence, in the place of ML technique, ANN is used as probability estimator with in the HMM structure, with out changing much of pure HMM system. (Claudio and Lucio, 1999) And this is the point that change the pure HMM approach to the hybrid ANN/HMM model as a new paradigm. In this model the MLP of the

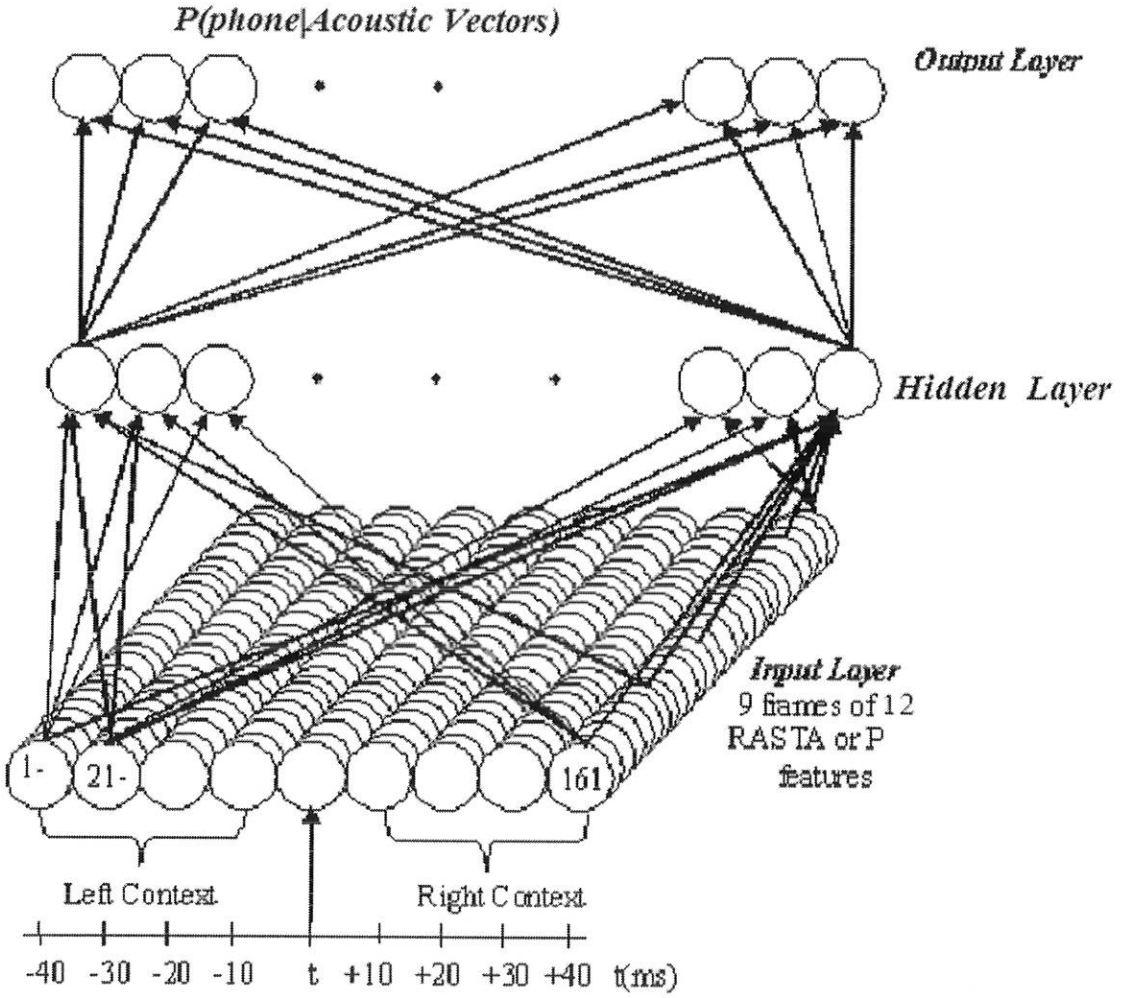


Figure 7 A Neural network used to estimate phone state probabilities (Modified from Jurafaky et al 2000, p. 270)

Figure 7 generally illustrates architecture of a neural network that is used for speech recognition systems with nine features at 10ms time over lap is fed to an input raw. In such cases the input layer will have 9 rows and 12 columns. Because each frame is a vector of 12 PLP features. The number of nodes at the output layer is dependant on the number of speech units, phones used in the system. There is no guide mark to determine the number of nodes at

the hidden layer. Some literatures suggests 200 and more. (Hosom et al, 1999)

As shown in figure 7, nine frames, for instance, are given for the input of the MLP: four consecutive frames before, four frames after, and one frame at time *t*. Then the MLP will have one output for each phone by restricting the sum of all the output units to one. This helps to calculate the state probability, q_j, of a state j given an observation vector o_t, $p(q_j|o_t)$.

By applying Bayes rule the observation likelihood we need for the HMM $p(o_t|q_j)$ can be computed as:

$$p(q_j|o_t) = \frac{p(o_t|q_j)p(q_j)}{p(o_t)}$$

Rearranging this equation gives:

$$\frac{p(o_t|q_j)}{p(o_t)} = \frac{p(q_j|o_t)}{p(q_j)}$$

Here probability of the observation $p(o_t)$ is a constant during recognition (Jurafsky et al, 2000), hence we can say that:

$$p(o_t|q_j) \propto \frac{p(q_j|o_t)}{p(q_j)}$$

3. *Signal Processing*

As mentioned in 2.2.2 the first element of all ASR systems is signal processing, commonly known as front end signal processing. In this phase, the speech waveform is converted to some type of parametric representation for further analysis and processing (Rabiner and Juang, 1993).

And, though there are many other types of possibilities, the short time spectral envelope is used as the most common important parametric representation of speech. By this method, the feature vectors used to characterize the spectral properties of the input speech are derived.

Generally, the natural speech wave, which is continuously varying over the time, is sampled at a fixed sampling rate, commonly 8000, 11025, 16000, 22050, 44100, etc. samples per second, and converted into a sequence of parameter vectors at a certain frame rate of mostly 10ms, 15ms, 20ms.

There are two dominant methods of spectral analysis: Linear Predictive Coding (LPC) and Mel-Frequency Cepstral Analysis[1]. LPC assumes that the speech signal is produced by the glottis and this speech can be characterized by its intensity (loudness) and

[1] Mathematical details are omitted here; the interested reader is referred to e.g. Rabiner and Juang (1993).

frequency, which determines the pitch of the sound. The vocal tract, that is the combination of the throat and the mouth, forms a tube, which is characterized by its resonance, called formants. By estimating the formants, removing their effects from the speech signal intensity and frequency of the remaining speech signal, LPC analyzes the speech signal frames (Wiggers, 2001). The basic intention of LPC is to determine the formants from the speech signal, which is done by a difference equation, called a linear predictor that expresses each sample of the signal as a linear combination of previous samples. The coefficients of the difference equation, the prediction coefficients, characterize the formants (ibid).

On the other hand, the Mel-frequency cepstral analysis deals with power spectrum of a speech signal which describes the frequency content of the signal over time. And this is done by a known mathematical function called Discrete Fourier Transform (DFT), which computes the frequency information of the equivalent time domain signal (Ibid).

However, as a speech signal contains only real point amplitude values, a real-point Fast Fourier Transform (FFT) will be performed for increased efficiency. The resulting output contains both the magnitude and phase information of the original time domain signal. Here a different scale called Mel-scale is employed

instead of a linear scale (ibid). Mel frequency cepstral coefficient MFCC features along with their derivatives are now accepted as the standard front-ends in ASR systems (Gu 2001).

4. *Language and Acoustic Modeling*

A general equation, $\hat{W} = \arg\max_{W} P(W|O)$ is used for all speech recognizers, where $\hat{W}$ is a word or sentence with maximum argument value of probability of the word or sentence given an observation acoustic signal, $P(W|O)$

By applying the Bay's rule $P(W|O)$ can be rewritten as:

$$P(W|O) = \frac{P(O|W) * P(W)}{P(O)},$$

From this equation it is possible to observe that ***P(O | W)*** is the acoustic model and ***P(W)*** is the language model used in an ASR system. This shows us that an ASR system has two fundamental components: the acoustic model and the language model (since ***P(O)*** is a constant, as noted above).

4.1.1 Language Modeling

The language mode provides the estimates of probabilities of word sequence in a given recognition task. The most common way of doing this task is by N-gram modeling. In this language modeling system there is a

logical assumption that the probability of a word sequence W can be reasonably computed as:

$$P(W_1^n) = P(w_1)P(w_2 \mid w_1)P(w3 \mid w_1^2)...P(wn \mid w_1^{n-1})$$
$$= \prod_{k=2}^{n} P(w_k \mid w_1^{k-1})$$

But the above equation is difficult to solve and there is no easy way to compute the probability of a word given a long sequence of preceding words. This problem is solved by making a useful simplification, that is, the probability of the word given the single previous word (Jurafsky and Martin, 2000). This simplification is known as the bigram model. Then the equation becomes:

$$P(w_1^n) \cong P(w_n \mid w_{n-1})$$

In most cases, the language model P(w) is estimated from a given (large) text corpus using a simple relative frequency approach (Zegaye, 2003)

$$P(W_i / W_{i-1}.....W_{i-N+1}) = \frac{F(W_i, W_{i-1},.....W_{i-N+1})}{F(W_{i-1},.....W_{i-N+1})}$$

Where, F is the number of occurrences of the string in its argument in the given training corpus.

5. *Acoustic Modeling and Decoding*

As discussed earlier the speech signal has been transformed into a parameterized form. Then it must be recognized, or decoded, and turned into the underlying sequence of symbols. This decoding process requires patterns or models against which unknown utterances can be compared (Odelly, 1995).

The acoustic model, $P(O|W)$, computes the probability that the speech data was observed for a given word sequence. Ideally, the model is constructed collecting many examples of words (W) and collecting the statistics of the corresponding vector sequences. For large vocabulary systems this is impractical. Hence, word sequences are decomposed into phonemes (Young, 1996), because phonemes of a language mostly are limited in number.

The direct estimation of the joint conditional probability p from instances of spoken words due to the dimensionality of the observation sequence O is also difficult. Thus a parametric model such as a Markov model can be assumed and the estimation from data will be possible. This is because the problem of estimating the class conditional observation densities $P(o|w_i)$ is replaced by the much simpler problem of estimating the Markov model parameters. Thus in place of W_i, HMMs of the base units like phonemes or words will be used (Zegaye, 2003). A

better alternative is using Neural Net as observation probability estimator. And this is the method used in the ANN/HMM hybrid model.

Once the language and the acoustic models are defined the probabilities for word sequences are generated as a product of the acoustic and language model probabilities and the process is known as decoding or searching. In this stage, "a directory of word pronunciations and a language model (probabilistic grammar) are taken and a Viterbi or A* decoder are used to find the sequence of words which has highest probability given the acoustic events." (Jurafsky and Martin, 2000). These two algorithms search an optimal sequence of states q_i to define or approximate the recognized word. The Viterbi algorithm uses a bigram language model and synthesizes the recognized word assuming that the best state at a given point of time is best for the entire sequence of observation. Hence another most common alternative algorithm is the A* (Stack) decoder [2] (ibid). The A* decoding algorithm iteratively chooses the best prefix-so-far, computes all the possible next state for that prefix, and adds this extended word to the queue. Then by using an evaluation function called A* evaluation function, it selects the best optimal word.

Generally an ASR system has a front end in which the natural speech wave is digitized and parameterized for the recognizer. The

[2] Anyone, interested in a detailed description of the algorithms, can consult Jurafsky and Martin (2000).

recognizer has a Neural net to train on these digitized and parameterized data. After training, the neural net produces the estimation of probabilities of observations for the HMM states. The HMM uses these probabilities and the language model to compute the probability of a sequence of symbols, given the sequence of observation. And finally, the recognizer uses decoders to generate the recognized symbols as output.

Figure 8 summarizes the over all recognition processes performed by Hybrid HMM/ANN based recognizer. (ibid)

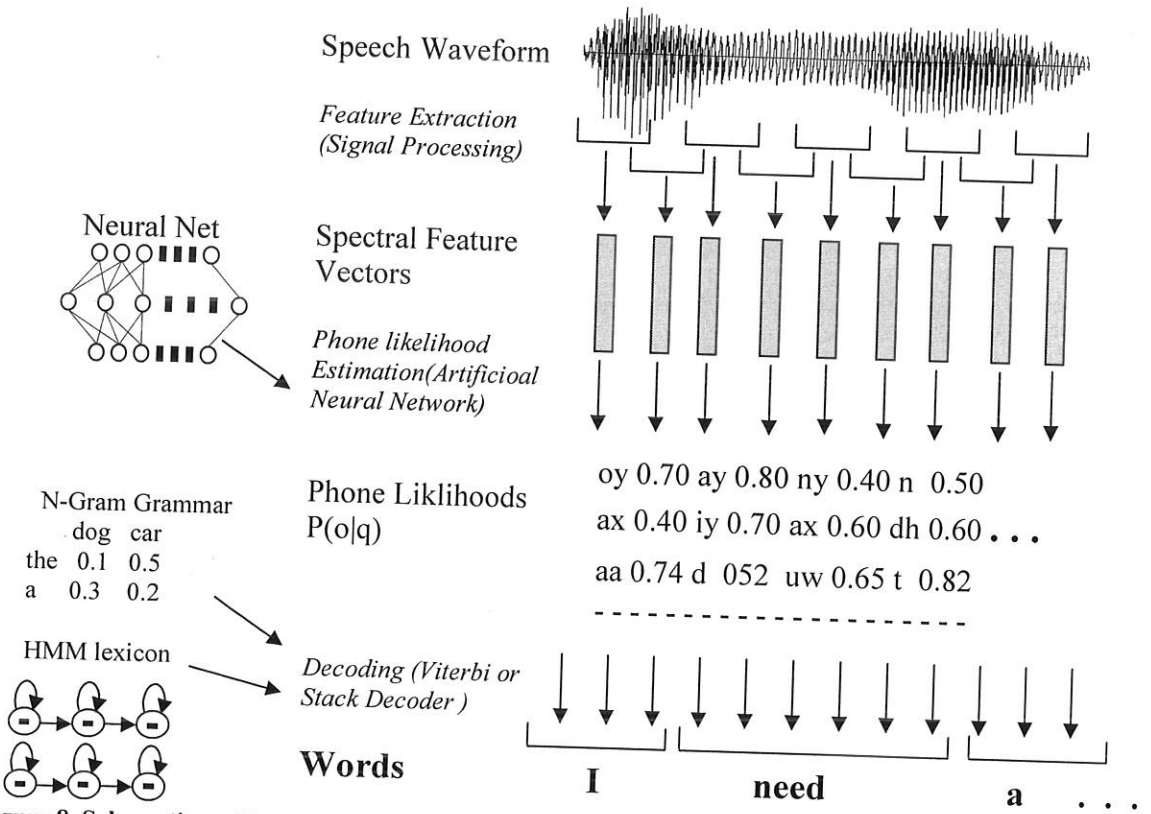


Figure 8 Schematic architecture for a (simplified) speech recognizer

Chapter IV An Amharic ASR System and the CSLU Toolkit

1. Introduction

The CSLU toolkit is a complex toolkit which is mainly designed not only for speech recognition, but also generally for both research and educational purposes in the area of speech and human-computer interactions. It is jointly develop and maintained by the Center of Speech Language Understanding, a research center at the Oregon Graduate Institute (OGI) of Science and Technology in Portland, Oregon (USA) and the Center for Spoken Language Research (CSLR) of the University of Colorado, which branched off from CSLU in 1998 (Dybkjær et al, 2001).

The CSLU toolkit, which is available free of charge for educational, research, personal, and evaluation purposes under a license agreement, supports core technologies for speech recognition and speech synthesis, plus a graphical based rapid application development (RAD) environment for developing spoken dialogue systems (McTear, 1999).

Regarding speech recognition, the CSLU Toolkit supports the development of HMM or ANN/HMM hybrid based speech recognition systems. For this purpose it has many different modules or tools

interacting with each other in an environment called CSLU-HMM (Hosom et al, 2000).

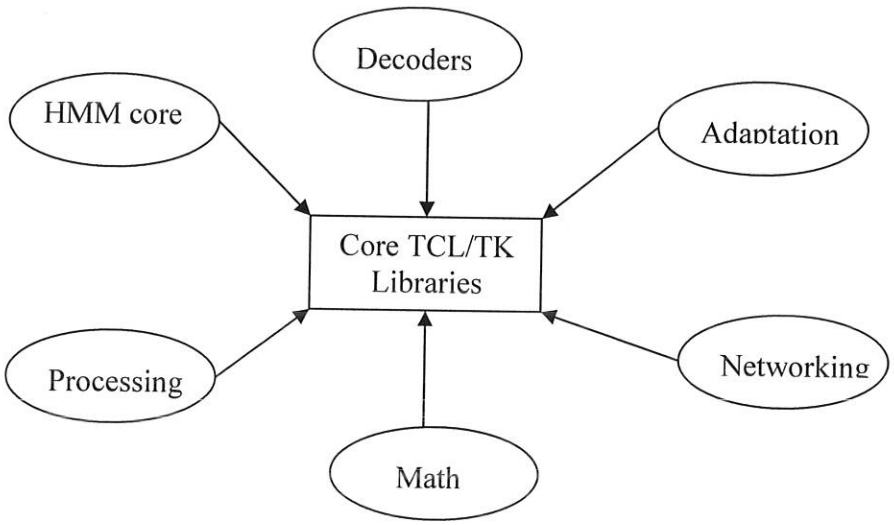


Figure 9. Software architecture of the CSLU-HMM environment

2. *Data Preparation Tools*

It is clear enough that before training the models and building the recognizer, the required speech data should be prepared and made available. Plus, the data should be preprocessed first in accordance to the requirements of the various functions of the toolkit. Having this common sense, the data preparation component of the Toolkit is described below.

The CSLU Toolkit installation provides a friendly graphic based interface or application to prepare speech data. It is speech view window. The speech view window is used to record, display, save and edit speech signals in their wave format. Speech view has also spectrogram and some

other speech wave related data like pitch and energy counters, neural net outputs and phonetic labels.

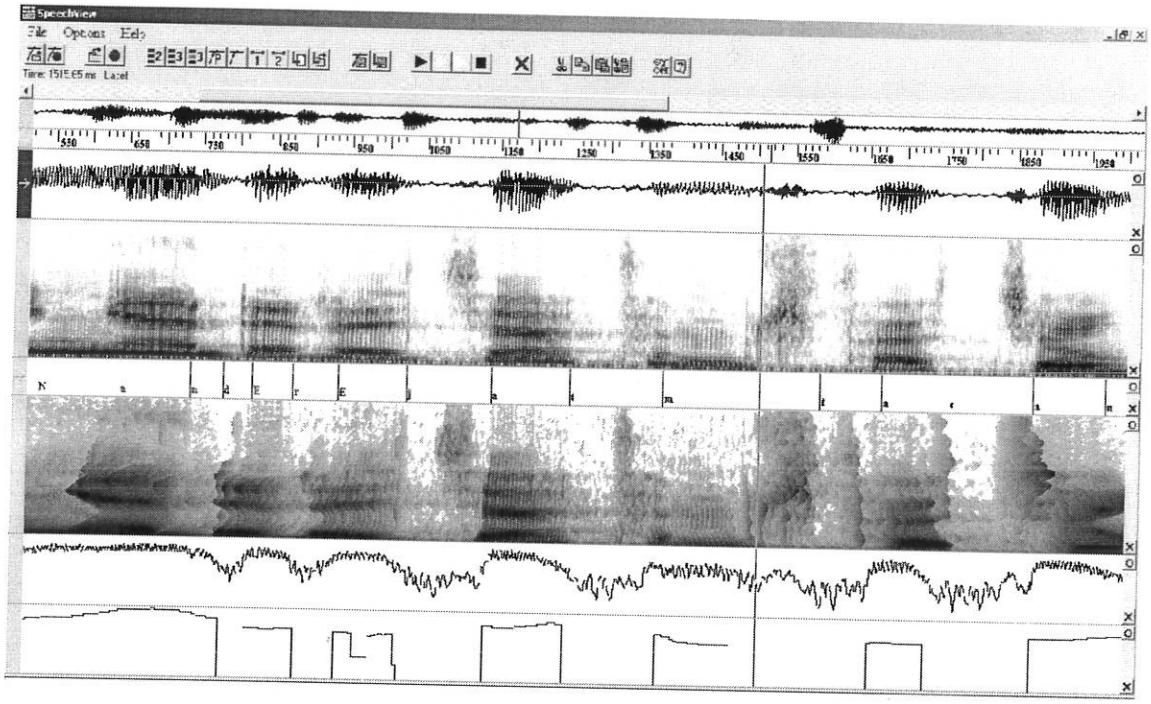


Figure 10. Speech view window of the toolkit

With the help of all these tools, one can collect and prepare speech data in an easy way for training a recognizer. The annotation process of the speech waveform, which is the most tedious and difficult process in the development of speech recognition systems, can be done in a friendly environment at different levels of transcription (phonetic, word or other level).

The speech view window can also be used to play back a selected part of a wave or the whole waves just by clicking on the play buttons differentiated by colors. It also provides the facilities for cutting parts of

speech or silences as needed. The label window of the speech view facilitates the basic functionalities of annotation. The label window can be divided into cells aligned with the relevant segment of other displays. The cell span can be varied simply by dragging the mouse over their edges. The phonetic or word transcription can be supplied just by typing the needed labels on the cell. The figure shown above illustrates the scenario. All these facilities make the speech data collection and preparation process relatively easy and comfortable.

3. *Data Processing Tool*

Once the CSLU toolkit is installed, different commands can be executed through DOS, cmd.exe of Windows or by its own DOS based shell program called CSLUsh. For data processing there is no graphical interface, instead there are many tcl script commands to be forwarded on DOS TCLsh, CSLUsh or unix interface window.

In general, to construct an HMM/ANN hybrid model speech recognition system we need to specify the phonetic categories that the network will recognize and find many samples of each of these categories in the speech data (Cole et al, 1999). Finally we need to train a network to recognize the specified categories and evaluate its performance using a test data set. Here Cole et al (1999) suggest the following four steps to follow.

- Divide the waveform into frames, where each frame is a small segment of speech that contains an equal number of waveform samples
- Compute features for each frame
- Classify the features in each frame into phonetic based categories using a neural network
- Use the matrix of probabilities and a set of pronunciation models to determine the most likely word(s)

In the CSLU toolkit there are different TCL script commands to implement all these steps. The following list describes the main script commands and important description files used in implementation of the ANN/HMM hybrid model based ASR systems. The description files are files, which have specific format and are used to describe the recognizer and the data that will be selected for training. They could be created manually or automatically by some special commands.

Table 2. List of tcl commands used in speech recognition

1. Description Files

- corpora file — This file is used to describe the format and location of the file in the corpus on which the recognizer trains. It contains a master list of each corpus and it is created manually.

- cull file — This file contains the list of wave files from the corpus that will be used for in-house testing purposes, usually 5% of the whole corpus. It is optional and can be created by the script cull5.tcl
- vocab file — This file is a vocabulary file and contains the pronunciation and grammar of the recognizer. It is created manually.
- parts file — 'parts' file specifies how many parts to split each phoneme into, and what context groupings to use.

2. **Tcl Command files**

- categories.tcl — Used to automate the process of determining categories
- find_files.tcl — Finds files for training, development, and testing
- gen_catfiles.tcl — Create time aligned categories from text transcriptions or from time aligned transcriptions
- revise_desc.tcl — Checks that all phone categories have enough samples for training
- hscript .exe — Create other files that will be used in training and recognition

- pickframes.tcl Selects samples to train on
- genvec.tcl Generates features for each frame to be trained on and creats vector filesin a temporary folder.
- checkvec.exe Checks the validity of data in the vector files
- nntrain.exe Train the network on the vector file generated by genvec.tcl command.
- Find_best.tcl Finds the best iteration of the network using the set of development files
- Browse.tcl Evaluates the errors so that it may give clues about necessary revisions to the recognizer

In the process of speech recognition development, the most time taking and boring task is annotating speech data. Of course there should be enough annotated speech data to develop better recognizer and to get good performance. Unfortunately it is impossible to annotate the whole collected speech data, especially if the collected speech corpus is very big.

To solve this problem, the toolkit implements a fantastic feature known as forced alignment. Forced alignment is an automatic annotating method by using a previously trained network. The previously trained network takes word transcription text of a speech file with the

corresponding speech wave file and produces a time aligned annotation of the speech data.

- Finally the network will be retrained on this automatically annotated data including the previously used data. After this a third network will be created by using the forward backward method; for this purpose hmm_emded.tcl, hnncheckveck.exe and hnntrain.exe functions are used. For more detailed information one can consult Cole et al (1999).

This process is further automated by Ben Serridge at the Tlatoa, a speech technology group[3] at Universidad de Las Americas in Puebla (Kirschning, 2001). In this automation all the commands and functions mentioned above are used in such a way that one or more functions are called and interact to serve their purpose. The functions used by Tlatoa are listed in their technical documentation.

[3] Available at: http://info.pue.udlap.mx/~sistemas/tlatoa/howto/techdoc.html

Chapter V Experimentation

1. Introduction

The attempt of this research is to test the application of ANN/HMM hybrid model based speech recognizer for Amharic language capable of recognizing Amharic speech. The recognizer is designed to use large vocabulary, to recognize continuous speech and is speaker independent. This was implemented using phonemes as base unit. Though due to time constraints the system is trained on limited words, the system's vocabulary is not limited in any way to some fixed set of words. By modifying the language model and adding the new word to the pronunciation dictionary, it is possible to incorporate a new vocabulary.

For this research the ANN/HMM hybrid model approach is implemented and this is the difference from the work done by Zegaye (2003). Zegaye used a pure HMM and reached 76.2 word and 26.06% sentence level accuracy. Here the intention of using ANN/HMM hybrid model approach is to find another alternative for better performance. Because, literature shows that better performance can be achieved by this approach.

The development process was performed using the tools in the CSLU toolkit installed on the Microsoft Windows 2000 platform. Various preprocessing programs and script editors were also used. The

discussion of the experiments is presented in accordance to the steps that should be followed while building such a speech recognizer.

2. *Data Preparation*

It is understood that for the development of an ASR system, the required input data are speech data, transcription of the speech data and a large text document. Hence the required data should be prepared and made available before training the models and building the recognizer. In effect, the expected major deliverables of data preparation are pronunciation dictionary, speech files, and annotation (label) files and text transcription of the speech files.

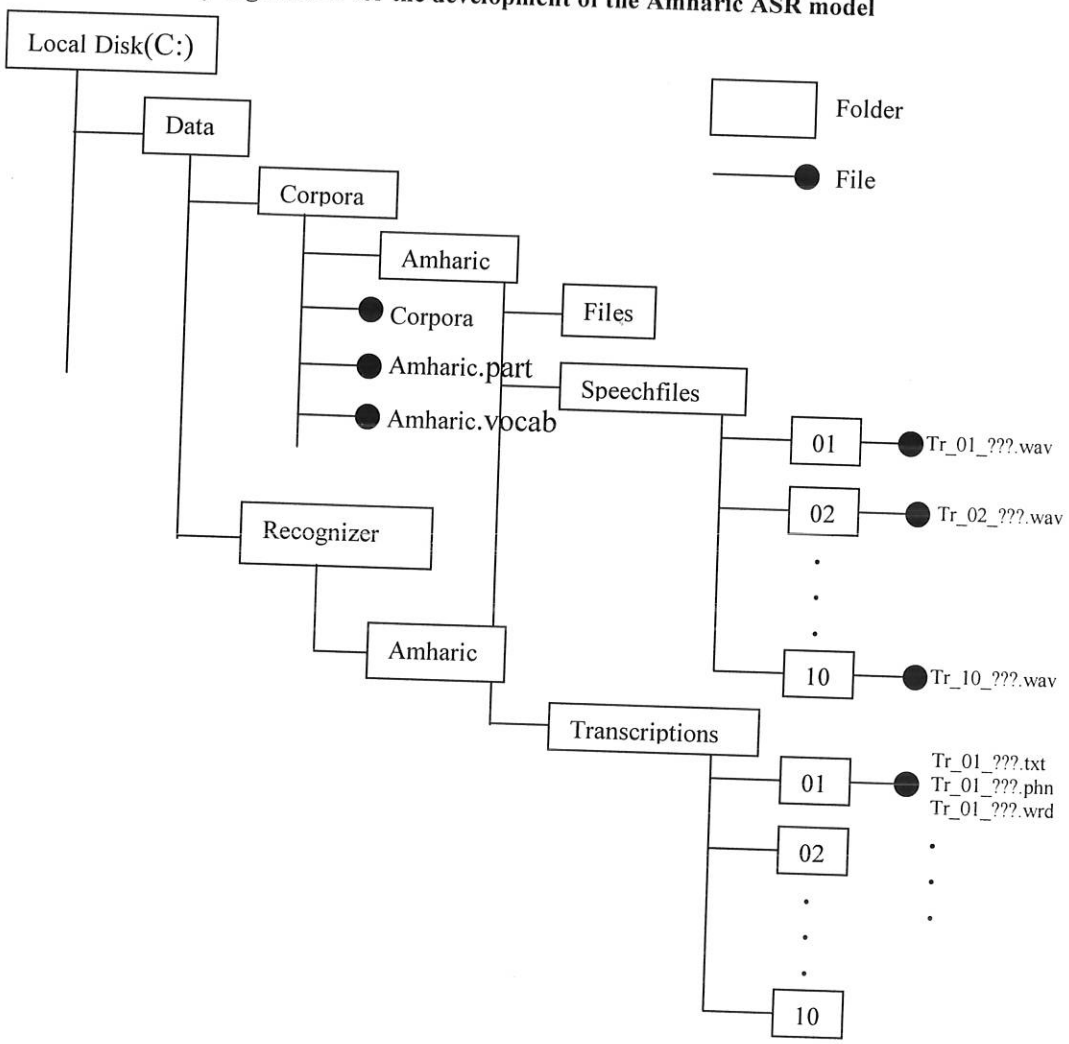
Since there does not exist a standard corpus already developed for the Amharic language, the experiment had to go to the extent of preparing and processing the corpus using the speech view tool of the CSLU toolkit. Still, this research has got an advantage of using 50 peoples recorded data at 16000 sampling rate from Solomon (2001). Each speaker has been recorded for 100 different sentences in a separate folder with a file for each sentence.

Considering the time constraint, only the first ten speakers on the first ten spoken sentences are annotated phonetically and **.phn** files are generated for each wave file. The scripts of each wave file are saved in text format with **.txt** extension file. Time aligned word label transcriptions

that are generated automatically, are saved in **.wrd** file extension. All folders containing wave files are saved in C:\data\corpora\Amharic\speechfiles, while all folders containing transcription and annotation of each speaker are saved under C:\data\corpora\Amharic\transcriptions. The folders of each speaker are named 01, 02, 03, . . . 50, while file names are organized as tr_speeker folder number_sentence numbers. For example, for sentences of the fourth speaker whose wave file is saved under folder name 04, the file names are tr_04_000, tr_04_001, tr_04_002, . . . tr_04_099.

Since the CSLU toolkit needs a consistent organization of directories and files, the structure shown in the figure below is used for this prototype. After the development of the shown file and directory organization, the procedure given on the Tlotoa technical documentation webpage is strictly followed.

Table 3. File and Directory orgaization for the development of the Amharic ASR model



1. Having speech files and their phonetic level, there is a need to generate the label file on word level. Hence at the corpora directory a batch file **mk.bat** which contains **mk_files_file.tcl** for each speaker who needs to be labeled is created.

2. Files with `.file` extensions which contain the address of the associated files for each speaker are created by using **mk.bat**

3. To generate time aligned word transcriptions, another batch file `txt2.bat` is created. **`txt2.bat`** contains the `tcl` **`txt2wrd.tcl`** command for each speaker.

4. By executing the file **`txt2.bat`**, the time aligned word transcriptions of each speaker are generated automatically.

5. The word transcriptions are edited with the help of the speechview tool.

6. Another batch file **`trm.bat`**, which contains the `tcl` **`wrd_trim_wav.tcl`** command for each speaker is created and executed to trim long silences at the beginning and end of the wave file and to adjust transcriptions accordingly.

7. Since the time alignment of phonetic description is disturbed because of long silences removal, the phonetic label is generated automatically from timed aligned word transcriptions by using the `tcl` **`wrd2phn.tcl`** command.

8. The speechview tool is used to edit the newly generated phonetic labels once again.

9. After this the boundaries of the word transcriptions are adjusted by the `tcl` **`adjust_wrd_boundaries.tcl`** command.

10. Based on the pronunciation of each word to be used in this research and parts of the phonemes discovered earlier, the vocabulary file

Amharic.vocab and category file **Amharic.part** are created in a text editor (WordPad)

11. The recognizer is created as: **make_recognizer.tcl** -name amharic -corpus amharic -sampling_rate 16000

12. Taking the part file as an input, cat files are generated on which the recognizer trains as: **generate_cat_files.tcl** -name amharic -corpus amharic -files amharic.train.amharic.files

13. Frame vectors are generated by **gen_data.tcl.** Here a file amharic.train.vec containing the feature vectors for each category is created. The command looks like: **gen_data.tcl** -name amharic -corpus amharic

14. Now everything is ready to train the network. By using **train_and_test_nets.tcl**, the network is trained and tested automatically by typing the command: **train_and_test_nets.tcl** -name amharic -corpus Amharic

15. At this stage the network is trained and tested. The test results are collected and for further testing purposes the following command is used: **find_best.tcl** nnet Amharic.test.Amharic.files Amharic.vocab Amharic.train.olddesc hand_labels.summary -b 15 -g 10

3. Test Results

In the CSLU toolkit, the evaluation step is very simple. A single command, **`find_best.tcl`** does that. In this research the recognizer is evaluated on ten peoples' speech for two sentences each. Then the results are as shown in the table below.

Itr	#Snt	#Words	Sub%	Ins%	Del%	WrdAcc%	SntCorr
15	20	236	13.62%	4.89%	5.83%	75.66%	42.31%
16	20	236	13.62%	5.83%	5.83%	74.72%	42.31%
17	20	236	13.62%	4.89%	6.83%	74.67%	41.72%
18	20	236	14.61%	4.89%	5.83%	74.67%	42.31%
19	20	236	15.56%	3.89%	4.89%	75.66%	41.72%
20	20	236	11.67%	5.79%	4.89%	77.65%	42.90%
21	20	236	11.67%	5.83%	4.89%	77.61%	42.90%
22	20	236	14.61%	5.83%	5.83%	73.73%	41.13%
23	20	236	13.62%	4.89%	4.89%	76.61%	42.90%
24	20	236	13.62%	2.93%	5.79%	77.66%	42.90%
25	20	236	14.61%	2.93%	4.89%	77.57%	42.31%
26	20	236	14.61%	4.89%	4.89%	75.62%	42.31%
27	20	236	15.56%	3.89%	4.89%	75.66%	42.31%
28	20	236	12.66%	3.89%	4.89%	78.56%	44.07%
29	20	236	12.66%	5.83%	4.89%	76.62%	42.31%
30	20	236	12.66%	4.89%	4.89%	77.56%	42.90%

Best results (78.56%, 44.07%) with network nnet.28

Where Itr, #Snt, #Words, Sub%, Ins%, Del%, WrdAcc%, and SntCorr in the table above show us the iteration number, number of sentences used for evaluation, number of words used for evaluation, substitutions, deletions, insertions, word accuracy, and percentage of correct sentences respectively. The best results (78.56% word level accuracy and 44.07% sentence level accuracy) are obtained on the 28th iteration.

Here all the speakers are found in the training data. When the same recognizer is tested for another two speakers who were not included in the training data with two sentences each, the recognition rate degrades. The results are shown below.

Itr	#Snt	#Words	Sub%	Ins%	Del%	WrdAcc%	SntCorr
15	20	218	16.34%	5.87%	7.00%	70.79%	35.27%
16	20	218	16.34%	7.00%	7.00%	69.65%	35.17%
17	20	218	16.34%	5.87%	8.20%	69.59%	33.79%
18	20	218	17.53%	5.87%	7.00%	69.60%	34.27%
19	20	218	18.68%	4.66%	5.87%	70.80%	33.79%
20	20	218	14.00%	6.93%	5.87%	73.20%	36.75%
21	20	218	14.00%	7.00%	5.87%	73.13%	35.35%
22	20	218	17.53%	7.00%	7.00%	68.46%	33.62%
23	20	218	16.34%	5.87%	5.87%	71.92%	37.75%
24	20	218	16.34%	3.52%	6.95%	73.19%	34.75%
25	20	218	17.53%	3.52%	5.87%	73.08%	34.27%
26	20	218	17.53%	5.87%	5.87%	70.73%	34.27%
27	20	218	18.68%	4.66%	5.87%	70.80%	34.27%
28	20	218	15.19%	4.66%	5.87%	74.28%	43.70%
29	20	218	15.19%	7.00%	5.87%	71.94%	35.27%
30	20	218	15.19%	5.87%	5.87%	73.07%	35.64%

Best results (74.28%, 43.70%) with network nnet.28

The word accuracy is reduced by 4.28% while the sentence level recognition rate is reduced by 4.37%.

The result in this research (21.44% word level error rate and 55.93% sentence level error rate) has a 2.36% decrease in word error rate and 18.01% decrease in sentence error rate compared to Zegaye (2003), who reported 23.80% word and 73.94 sentence error rates. The relative error reduction is thus 9.92% ($\frac{23.8-21.44}{23.8}\times 100\%$) at the word level and 24.36% ($\frac{73.94-55.93}{73.94}\times 100\%$) at the sentence level.

Chapter VI Conclusion and Recommendations

4. Conclusion

Given the circumstances and the limitations of time and experience, the results obtained in this research are promising. As the main goal was to see the possibilities of the Hybrid ANN/HMM approach for Amharic speech recognition, the results of the experiments indicate that a better recognizer can be developed with further optimization efforts.

The CSLU toolkit also can be considered as good toolkit to develop Hybrid ANN/HMM based speech recognizers. But it needs some revisions to fix the implementation problems of the toolkit in the Windows environment.

There were problems to download the Toolkit Installer completely. Even after installation the integration problem of the toolkit with the Windows operating system has consumed most of the schedule. This makes the researcher unable to prepare more training data and to improve the recognizer. If that were not the case, the result of the evaluation would have been improved further.

The data used in this research was labeled manually by the researcher himself. But the task needs experts to identify each phone in a better way. The quantity and the quality of the data used have significant influence on development and performance of the speech recognition system.

5. Recommendations

In all speech recognition development research, most of the time schedule is wasted on the data preparation phase. Hence to reduce this time it is good idea to develop a nation wide speech corpus. This corpus should cover most sociolinguistic aspects of the language. For English and other international languages there are such corpora.

A full-fledged functional ASR system for Amharic is needed for the application indicated in chapter 1, section 9. For this task, by its nature, a group from electrical engineering in the signal processing area, linguists, and computer and information scientists should be constituted and a detailed survey on the subject should be carried out.

As literature indicates and as this research confirms, the application of artificial neural networks to speech recognition systems gives a better performance. Hence further efforts should be undertaken with more data to utilize the approach to the limit.

The obtained results from this study may be far from the ideal intentions, but integrating some other supportive tools like spellchecking, the gap can be further diminished.

In addition to the preparation of speech corpus, the preparation of a pronunciation dictionary of the language should be focused, especially by linguists. The existence of such a dictionary might be of much service in the development of ASR systems.

Reference

1. ባዬ ይማም፣ ፈደል አንደኛና የአትዮጵያ ቋንቋዎችና የሥነ ጽሑፍ መጽሔት ቁጥር 7፣ (1 – 32)። 1997

2. Almeida, Luis B., Martins, Ciro and Neto, Joäo P.1(996) ***An Incremental Speaker Adaptation Technique For Hybrid HMM-MLP Recognizer,*** Available at: http://www.asel.udel.edu/icslp/cdrom/vol3/510/a510.pdf

3. Beaufays, F., Bourlard, H., Franco, H. and Morgan, N. (2001) ***Neural Networks in Automatic Speech Recognition.*** Available at: ftp://ftp.idiap.ch/pub/reports/2001/rr01-09.pdf

4. Bilmes, Jeff A. and Kirchhoff, Katrin (1999) ***Dynamic Classifier Combination in Hybrid Speech Recognition Systems Using Utterance-Level Confidence Values.*** Available at: www.icsi.berkeley.edu/ftp/global/pub/speech/papers/icslp00-cmi.pdf

5. Bourlard, H. and Morgan, Nelson (1997) ***Hybrid HMM/ANN system for speech recognition: overview and New Research Directions,*** In International school on Neural Nets: Adaptive Processing of Information.

6. Bush, Mascia A. and Kopec, Gary E. (1985) ***Network Based Connected Digit Recognition Using Vector Quantization***, IEEE

Unpublished M. Sc. Thesis, Addis Ababa University: Faculty of Informatics, Addis Ababa

14. Laine Berhane. (1998) ***Text-To-Speech Synthesis Of The Amharic Language.*** M.Sc Thesis, Addis Ababa University: Technology Faculty, Addis Ababa

15. Lea, Wayne A. (1982) ***"Speech Recognition" In Encyclopedia Of Science & Technology***, 5th Ed. pp 875-879. New York: McGraw-Hill.

16. Lee, Kai-Fu (1989) ***Automatic Speech Recognition***, Boston: Kluwer Academic publishers

17. Lippman, Richard (1997) ***Speech Recognition by Machine and Human***, J. Speech Communication. Vol. 22 (1-15) Elsevier

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

19. Martha Yifiru (2003) ***Automatic Amharic Speech Recognition System To Command And Control Computers***, Unpublished M.Sc. Thesis, Addis Ababa University, Addis Ababa

20. McTear, Michael F. (1999) ***Using CSLU Toolkit for Practicals in spoken dialogues technology,*** School of Information and Software Engineering, University of Ulster, Ireland. Available at: http://cslu.cse.ogi.edu/toolkit/pubs/pdf/mctear_MATISSE_99.pdf

21. Moore, Roger K. (2002) ***A Review Of Large Vocabulary Continuous Automatic Speech Recognition***, IEEE Signal Processing Magazine, pp. 45-57

22. Nahm, Eric and Slater, Deborah (1997) ***"Speech Recognition"*** In The Ultimate Multimedia Handbook Edited by Jessica Keyes. New York: McGraw-Hill.

23. Odell, Jullian James (1995) ***The Use of Context in Large Vocabulary Speech Recognition***. Dissertation Submitted to the University of Cambridge: Queens College.

24. Oregon Graduate Institute of Science and Technology, Available at: http://cslu.cse.ogi.edu/tutordemos/nnet_training/tutorial.html

25. Rabiner, L.R., and B-H, Juang. (1993) ***Fundamentals of Speech Recognition***. Englewood Cliffs, New Jersey: Prentice Hall, Inc.

26. Reichl, W. and Ruske, R. (1996) ***Hybrid RBF-HMM System for Continuous Speech Recognition***, Munich University of Technology, Germany.

27. Rodman, R. D. (1999) ***"Speech Recognition By Machine: A Review" Reading In Speech Recognition***, Eds. Waibel, Alex and Kai-Fu Lee, California: Morgan Kaufmann.

28. Rottland, Jörg, Neukirchen, Christoph and Rigoll, Gerhard (1996) ***A New Hybrid System Based On MMI-Neural Networks For The RM Speech Recognition Task,***

Available at: www.mmk.ei.tum.de/~waf/publ/96/atlanta-96.pdf

29. Schalkwyk, J, P. Hosom, , E. Kaiser, K. Shobaki, (2000) ***CSLU-HMM: the CSLU Hidden Markov Modeling Environment.*** Available at: http://cslu.cse.ogi.edu/tutordemos/csluhmm/doc/csluhmm.ps

30. Solomon Berhanu (2001) ***Isolated Amharic Consonant-Vowel (CV) Syllable Recognition: An Experiment Using the Hidden Markov Model.*** Unpublished M. Sc. Thesis, Addis Ababa University: Faculty of Informatics, Addis Ababa

31. The Random House Dictionary of the English Language, Random House, New York, 1973

32. Tuerk, Andreas (2001) ***The State Based Mixture Of Experts HMM With Application To The Recognition Of Spontaneous Speech***, Ph.D. Dissertation, University of Cambridge

33. Wiggers, Paskal. (2001) ***Hidden Markov Models for Automatic Speech Recognition and their Multimodal Applications***. Delft University of Technology: The Netherlands.

34. Worku Alemu (1997) ***The Application of OCR Techniques to the Amharic Script. Unpublished M. Sc. Thesis, Addis Ababa University***: Faculty of Informatics, Addis Ababa.

35. Young, Steve et al.. 2000. ***The HTK Book***. Microsoft Corporation.

36. Zegaye Seifu (2003) ***HMM Based Large Vocabulary, Speaker Independent, Continuous Amharic Speech Recognizer***, Unpublished M.Sc. Thesis, Addis Ababa University, Addis Ababa.

37. Zegers, Pablo (1998) ***Speech Recognition Using Neural Networks***, Department of Electrical and Computer Engineering, the University of Arizona

38. Zue, V. and Cole, R. (1995) ***"Spoken Language Input Overview" Survey of the State Of Art in Human Language Technology***. Oregon Graduate Institute

Annexes

A_ 1 Parts and Average Duration of Phonemes

Phonim	Frequency	Average duration in ms	Part
ሯ	2	91.50000	3
ሸ	4	66.00000	1
ቈ	20	75.15000	2
ቱ	21	91.90476	3
ጅ	23	88.21739	2
ሽ	25	95.36000	3
ፅ	28	80.17857	2
ፐ	32	86.46875	2
ኝ	34	84.17647	2
ዝ	41	79.19512	2
ፍ	60	83.33333	2
ጭ	62	78.17742	2
ቅ	69	80.00000	2
ኤ	80	75.48750	2
ሀ	85	60.25882	1
ክ	106	81.60377	2
ኾ	133	105.79699	3
ግ	151	75.35099	2
ኡ	169	61.86982	1
ኢ	177	70.79096	1
ድ	182	63.13736	1
ብ	190	73.33684	1
ሿ	193	72.35233	1
ኢ	194	60.40722	1
ስ	197	86.83249	2
ር	225	61.23111	1
ሙ	244	63.31557	1
ል	282	72.19504	1
ይ	291	57.94158	1
ም	295	77.86102	2
ት	300	73.80667	1
ን	417	71.03357	1
አ	683	76.98389	2
ኦ	872	62.10894	1

A_ 2. The Amharic Phonemes (Modified from Zegaye's work)

Amharic Symbol	Etop font	In this paper	Amharic Word Example	
ማ	m	m	/mar/	'honey'
ተ	t	t	/tEmari/	'student'
ነ	n	n	/nIgus/	'king'
ለ	l	l	/lIb/	'heart'
በ	b	b	/bet/	'house'
ር	r	r	/kIrIkIr/	'debate'
ደ	y	y	/ayn/	'eye'
ወ	w	w	/wIyIyIt/	'discussion'
ሰ	s	s	/sEw/	'man'
ገ	g	g	/gEmEd/	'rope'
ዳ	d	d	/dabbo/	'bread'
ከ	k	k	/kErEmela/	'candy'
ቸ	^c	c	/ ?IgIr/	'problem'
ጠ	.t	T	/t'Ewat/	'morning'
ቀ	k'	q	/k'ElEm/	'paint'
ሀ	h	h	/hasab/	'idea'
ፋ	f	f	/flagot/	'interest'
ጀ	^g	j	/jEmErE/	'he started'
ኘ	~n	N	/tE ñ a/	'he slept'
ሸ	^s	S	/ šErErit/	'spider'
ፀ	.c, .s	x	/s'Ehai/	'sun'
ጨ	^C	J	/č'ErEk'a/	'moon'
ፐ	p	p	/polis/	'polis'
ጐ	uA	@	/g^{w}dEña/	'friend'
ዠ	^z	Z	/žIwažIwe/	'swing'
ዘ	z	z	/zEmEd/	'relative'
ጰ	.p	P	/p'ap'p'as/	'bishop'
አ	a	E	/ErE/	'exclamation'
ኡ	u	u	/udEt/	'circulation'
ኢ	i	i	/itIyop'iya /	'Ethiopia'
ኣ	A	a	/abat/	'father'
ኤ	E	e	/eli /	'tortoise'
እ	e	I	/Inat/	'mother'
ኦ	o	o	/oromo/	'oromo'

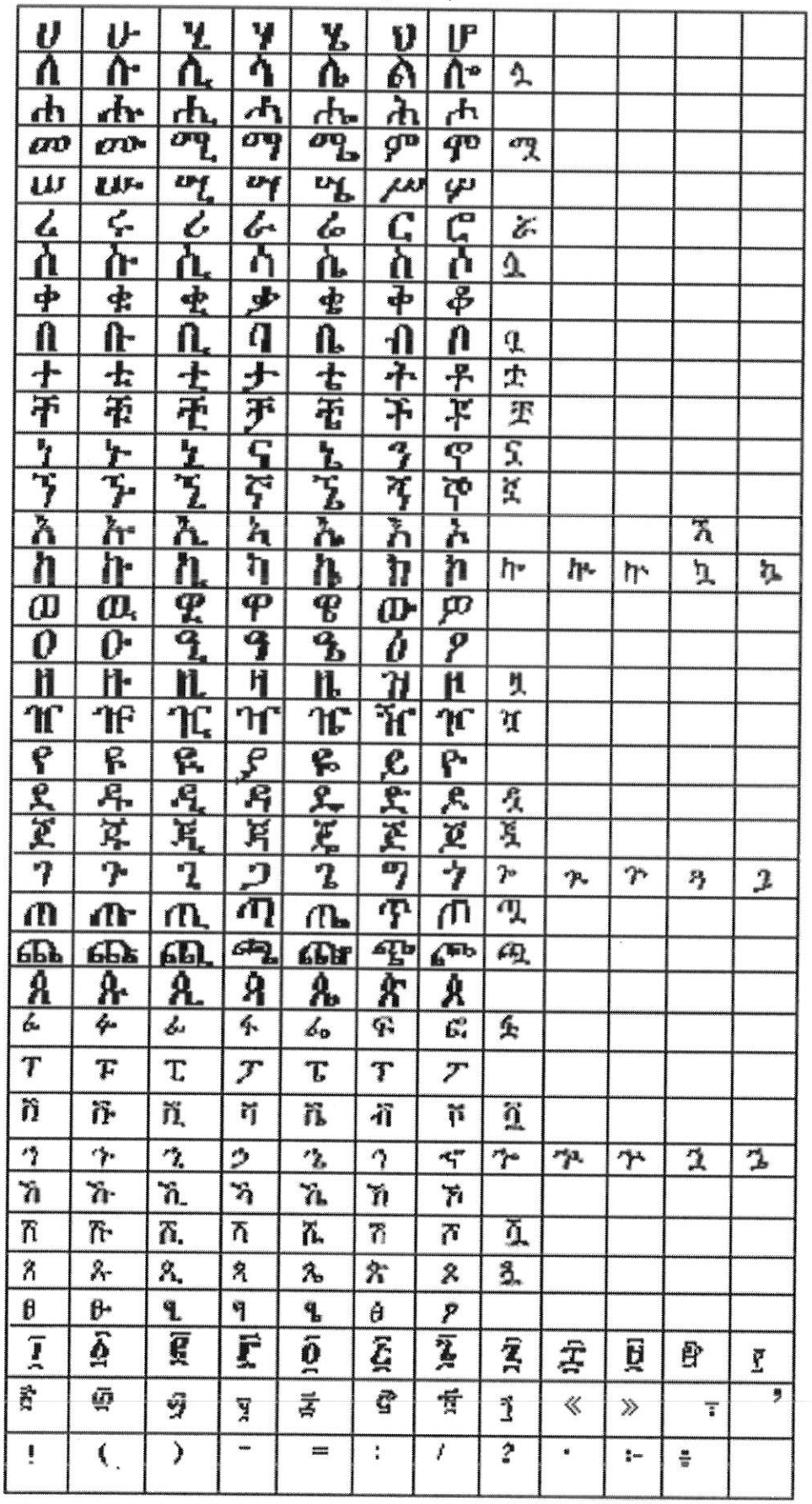
A_ 3. The content of the Vocabulary file Amharic.vocab

```
Ahun                {A h u n}                       ;
Banku               {B a n k u}                     ;
Gudayu              {G u d a y u}                   ;
Hod                 {H o d}                         ;
Husen               {H u s e n}                     ;

Etc. . .

yonas               {y o n a s}                     ;
ysaqe               {y s a q e}                     ;
yunvErstyt          {y u n v E r s t y t}           ;
zEmEn               {z E m E n}                     ;
zEgbwal             {z E g b w a l}                 ;
zEmEca              {z E m E c a}                   ;
zEndro              {z E n d r o}                   ;
zena                {z e n a}                       ;
zenwi               {z e n w i}                     ;

zmdna               {z m d n a}                     ;
zrzr                {z r z r}                       ;
zur                 {z u r}                         ;
pau                 {.pau}                          ;

separator    {.pau [.gar] .pau}                     ;

$qal    =
Ahun | Banku | Gudayu | Hod | Husen | INam | ISI | ITrEt | IbritacEwu
n | IdEhonm | IdErslacEwalEhu | IdgEt | Idiapa | Idme | Igr | IisayIa
s | Ijg | IlEt | ImiSu | ImidEnEq | ImigENu | ImigENutn | ImimElEkEta
cEw | ImnEt | ImnEtm | InEkbur | InErsum | InEsu | InEzih | Ina | Ina
ntEm | Inat | InawqalEn | InawqalEn | InclalEn | InclalEn | IndE | In
dEJErEsku | IndEgEba | IndEgEna | IndElelEbacEwm | IndElmadacEw | Ind

etc . . .

alu | yalutn | yamara | yan | yanIn | yanInIma | yandENa | yaqErEbEcb
acEw | yasgEnEzbalu | yasr | yastEdadralu | yaswErfat | yawETutko | y
awqtal | ybal | yclal | ydnEqacEw | yfElgal | ygEbal | ygEbwal | yhen
 | yhn | yhunna | yilEyayal | ymEgEnaNa | yonas | ysaqe | yunvErstyt
| zEmEn | zEgbwal | zEmEca | zEndro | zena | zenwi | zmdna | zrzr | z
ur;

$grammar = ([separator%%] < $qal [separator%%] > [separator%%]);
```

A_ 4. The Content of Amharic.part

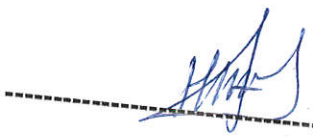
```
.pau  1 ;
d     1 ;
e     2 ;
E     1 ;
T     1 ;
s     2 ;
z     2 ;
n     1 ;
w     1 ;
I     1 ;
k     2 ;
f     2 ;
g     2 ;
r     1 ;
l     1 ;
p     2 ;
i     1 ;
u     1 ;
S     3 ;
a     2 ;
t     1 ;
o     1 ;
h     1 ;
b     1 ;
m     2 ;
N     2 ;
Q     2 ;
c     3 ;
j     2 ;
J     2 ;
Z     3 ;

$sil = .pau .gar;
$den = s Z S f; # dental
$bck = o u;     # back
$fnt = i e ;    # front
$cnt = I E a;   # center

map uc tc;
map uc kc;
```

A_ 5 Full Amharic Character Set (FidEl)



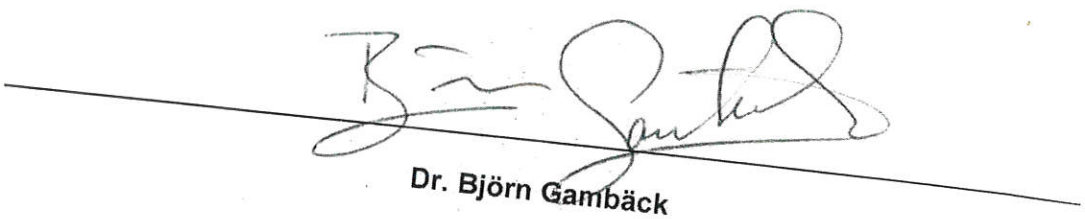
DECLARATION

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



Hussien Seid

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



Dr. Björn Gambäck