Large Vocabulary, Speaker Independent Continuous Speech Recognition for Ge`ez Language Using HMM

Assefa Mamo
HiLCoE, Computer Science Programme, Ethiopia
assefa.ma@yahoo.com

Wondowssen Mulugeta
HiLCoE, Ethiopia
School of Information Science, Addis Ababa University, Ethiopia
wondwossen.mulugeta@aau.edu.et

Abstract

A speech recognition system implements the task of automatically transcribing speech into text. This paper is concerned with automatic continuous speech recognition using trainable systems. The aim of this work is to build acoustic models for Ge`ez language. Ge'ez is the classical language of Ethiopia within the Semitic language family. Today, Ge'ez remains only as the main language used in the liturgy of the Ethiopian Orthodox Tewahedo Church. In order to accomplish the aforementioned objectives, speech data is recorded at a sampling rate of 16 KHz and the recorded speech is converted into Mel Frequency Cepstral Coefficient (MFCC). The corpus of 400 utterances has been developed that is used for training. A set of 100 additional sentences and phrases is uttered using the same speakers participated in the training phase and another set of 100 sentences and phrases is recorded using 5 different speakers that did not participate in the training phase to test and evaluate the system. An attempt is also made to build bigram language model with task grammar. Then, the acoustic model is built using the Hidden Markov Model approach. Tied-state triphones have been considered as the basic sub-word units that are extended from context independent phonemes in this study. Finally, the performance of the speech recognizer has been evaluated for trained speakers and the result was 64.21% word level correctness, 62.05% word accuracy, and 29% sentence level correctness. Moreover, 57.72% recognition accuracy, 62.19% word level correctness and 24% sentence level correctness has been obtained when the system is tested with speakers who had no involvement in the training phase. Based on the experimental results, the Pause between words can also affect recognition accuracy whether the set time is too short or too long.

Keywords: Speech Recognition; HMM; HMM based Speech Recognition; Language Modeling

1. Introduction

Speech is one of the oldest and the most widely and effectively used means of communication for humans. It plays a great role especially in human-machine interaction. Automatic speech recognition (ASR) or computer speech recognition is the decoding of the information conveyed by a speech signal into a set of characters. The resulting characters can be used to perform various tasks such as controlling a machine, accessing a database, or producing a written form of the input speech. Speech recognition can be classified into various types depending on how capable it is. These classifications are:

- Discrete (Isolated) vs. Continuous;
- Speaker-dependent vs. Speaker-independent;
- Large vocabulary vs. Small vocabulary speech recognition;
- Read or spontaneous speech;
- Noisy or Noise free speech.

People have extensively studied speech and often tried to build machines to handle it in an automatic way. During recent years, automatic speech recognition has started to emerge as a pragmatic and useful technology for various applications. During the last two decades, a remarkable advancement has been achieved in the field of ASR technology.

The ultimate goal of ASR is to make a real time system to achieve 100% accuracy in recognizing all words which are independent of vocabulary, speaker and environment characteristics. But the research is not close enough to meet the human ability [1]. There have been many works in ASR systems for technologically favored languages for more than 60 years in the globe. Unfortunately, research in local language speech recognition started only during the last two decades. Over the past few years, several natural language processing applications such as speech-recognition and Text-To-Speech (TTS) have been developed for local languages such as Amharic and Tigrigna.

Ge`ez Language

Ge`ez is an ancient South Semitic language that originated in Eritrea and the northern region of Ethiopia in the Horn of Africa. It later became the official language of the Kingdom of Aksum and Ethiopian imperial court [10]. Today, Ge'ez remains only as the main language used in the liturgy of the Ethiopian Orthodox Tewahedo Church, the Eritrean Orthodox Tewahedo Church, the Ethiopian Catholic Church, the Eritrean Catholic Church, and the Beta Israel Jewish community [10]. It is also taught as a second language in the traditional schools of the Ethiopian Orthodox Church [11, 12].

Ge'ez is fairly massive in size, with its 182 syllographs. There are essentially 26 main syllographs, all consonants, while the rest are essentially those with additional strokes and modifications added on to the main forms to indicate a vowel sound associated with it or to make aural adjustments in the basic consonant sound [13]. With a combination of every twenty six original alphabet with the seven vowel sound alphabets, a matrix of 26×7 size is produced.

As each language has its own set of phonemes from which sounds of words are generated, the classification of Ge`ez language`s consonants and vowels differ in their set of phonemes. Ge`ez language has its own characterizing phonetic, phonological and morphological properties.

Unlike many of the modern Ethiopian Semitic languages, Ge'ez language has the two pharyngeal consonants [ʔ](ዐ), [ḥ] (ሐ). In addition, the phoneme | **ኈ** | **ጎ** is treated as a laryngeal by the phonotatic rule of Geez. **ኀ** and **ኈ** are used in orthography of the word Tigrinya, even though **ኈ** does not exist in the spoken language [14].

Ge`ez consonants may also be subjects to gemination. Germination is a widely employed inflectional and derivational process in Ge`ez. For instance *('häqqäl' (ሐቀል) countryman, farmer* from *'häql' (ሐቀል) field* and *'t.äbbäbt' (ጠበብት) wise men from 't.äbib' (ጠቢብ).* All Ge`ez consonants can geminate with the exception it appears of the laryngeal. Ge`ez can also be pronounced with slight variations by putting the stress on different parts of a word to produce different sounds thereby giving more than one meaning to one and the same word [15]. To give an example *"yenägg`era" (ይነ ግሩ) they [f. pl] speak vs "yenägger`a" (ይነ ግሩ) he speaks to her*. Sometimes words are compromised with a certain style of pronunciation: high level, low level, acute, grave accent and extra short [16].

2. Related Work

Amharic speech recognition of isolated Consonant-Vowel syllable was developed using Hidden Markov Model Tool Kit (HTK) based on HMM with recognition accuracy of 87.68% and 72.75% for speaker dependent and independent systems, respectively [2]. In this system about 41 CV syllables of Amharic language from eight people (4 male and 4 female) with the age range of 20 to 33 years were used.

Amharic speech recognition using sub-word units was also examined for the purpose of comparative analysis using phonemes, CV syllable and tied-state triphones as the basic sub-word. About 170 word vocabulary recorded from 15 speakers (8 male and 7 female) in the age range of 20 to 30 for both speaker dependent and speaker-independent models to obtain

performance for the training data using tied-state triphone models (91.46% vs. 77.33%) and for the testing data of practically speaking (77.87% vs.75.33%) [3].

Amharic speech input interface to command and control Microsoft Word was also designed using HTK based on HMM-based, speaker independent, small vocabulary, and isolated word recognizers [4]. Fifty command words were used from 26 people to develop the prototype system. The system was tested using 18 randomly selected command words and it performed accurately in 16 commands. The comparative study in [2] was tried to compare two different toolkits: HTK and the MSSTATE Toolkit from Mississippi State for small vocabulary speaker dependent continuous Amharic speech recognizers using 50 sentences with ten words (the digits). The recognition system developed using the HTK has 82.5% word recognition accuracy whereas the system developed using MSSTATE has 79% word recognition accuracy.

Amharic speech recognition system was developed to solve Out-Of-Vocabulary (OOV) words problem using morphemes as dictionary and language model units. For Small vocabulary (5k) system of morphemes as lexical and language modeling units was better than words and an absolute improvement (in word recognition accuracy) of 11.57% had been obtained [5]. In addition, there are researches done [6, 7, 8, 9] for speech recognition and Text-To-Speech (TTS) of Amharic and Tigrigna languages.

To the best of our knowledge, there is no published work on the development of speech recognition for Ge'ez. Since Ge'ez has its own features such as its characteristics of phonetic and stress, it is not possible to make use of an ASR developed for other languages. Therefore, the purpose of this study is to explore the possibilities of developing speech recognition for large vocabulary of Ge`ez language that helps anyone who uses Ge`ez to exploit the opportunities of information technology for the application of speech recognition. It also contributions to the automation activity of preserving and transferring of knowledge and heritage accumulated in Ge`ez language to the next generation. The knowledge might be accumulated in written form or transferred orally from past generations. The oral knowledge could be captured and documented in written formats. There are many elites in the church who can pass their knowledge in a spoken form, which can be documented in a written form using ASR. One of the ultimate goals of speech recognition is to create a human-like listening typewriter or automated secretary. In addition, ASR is used as part of Automated spoken language translation system and thus it can be used as one part to translate Ge`ez language to another language.

3. Data Collection and Preparation in HTK

Text and speech data should be prepared and made available before training the models and building the recognizer. This data is divided into 2 subsets; namely, training set and test set. The training set is used to train the recognizer while the test set is used to test the performance of the recognizer. All preparation and pre-processing steps are discussed in the subsequent sections.

3.1 Text Corpus

The set of speech recordings must be based on a proper written set of sentences and phrases. Therefore, it is crucial to create a high quality written (text) set of the sentences and phrases before recording them [17]. This work aims to produce Ge`ez phrases and sentences that are phonetically rich and balanced. To fulfil this aim, 250 sentences and phrases with a length ranging from 2 to 29 words are selected from prayer book (የቅዳሴ መፅሀፍ) and history book (የአጤ ልብነ ድንግል ታሪክ) written in Ge`ez . These selected sentences have been manually checked for grammatical, spelling, and abbreviation errors and have been manually corrected. Long sentences are shortened to have structurally simple sentences in order to ease readability and pronunciation.

3.2 Speech Corpus

Speech corpus is an important requirement for developing any ASR system. It is simply a set of recorded sentences. No readily available speech corpus exists so it is needed to be recorded. The database used for this work is a corpus with phonetically rich and balanced sentences and phrases used for training and testing the acoustic model. Baum-Welch algorithm is used to train the HMM models. This database is divided into training dataset and test dataset. The speech corpus was recorded by 11 male and 4 female Ge`ez speakers who represent a wide range of age group (20-50). The speech corpus was recorded using Audacity recorder software version 2.1.3 at a sampling rate of 16 KHz with a WAV audio file format. During the recording sessions, each speaker is asked to utter sequentially 20 different sentences (totally 200 different sentences). Additionally each speaker uttered the same 20 different sentences for training purpose. Thus, 400 utterances are recorded in total. A set of 100 additional sentences and phrases is uttered using the same speakers who participated in the training phase and another set of 100 sentences and phrases is recorded using 5 different speakers that did not participate in the training phase. AS Audacity is also audio editor software. Each utterance of the speaker is checked directly for errors and the duration of the pause between the words is also amended that affect the performance.

A total of 1.22h read speech corpus is collected with a minimum speech time of 1.21s, maximum speech time of 25.72s and an average speech time of 7.98s. The technical details of speech corpus are shown in Table 1.

Table 1: Speech Corpus Technical Details

Criterion	*Training*	*Testing*
No. of utterances (sentences)	400	Two test, 100 for the same speaker in training, 100 for different speaker (untrained)
Number of Words	1248	316 (116 from training data and 200 out of training data)
Average No. of Words/Sentence	11	7
Min and Max No. of Words/Sentence	2 and 29	2 and 19
No. of Speakers	10	5 different speakers that is not participated in training phase, each 20 utterances (sentences), 10 speakers that is participated in training phase, each 10 utterances (sentences)
Speakers' Age (years)	20 – 50	20 – 50
Speakers' Gender	8 males and 2 females	11 males and 4 females
Sampling Rate (Hz)	16 kHz	16 kHz
No. of Bits	16	16
No. of Channels	1 (mono)	1 (mono)
Size of Raw Speech	102 MB	18.8 MB

3.3 Transcription Files

Many of the operations performed by HTK assume that the speech is divided into segments and each segment should have a name or a label. The set of labels associated with a speech file constitute a transcription. Since there is no hand labelled data to bootstrap a set of models, a flat-start scheme has been used instead. Phone level transcriptions are expected to be prepared out of the utterances. This is because, as the system is phoneme based and then a triphone, later during training examples of the phones are needed to adjust the parameters of these models. That is, to train a set of HMMs every file of training

data must have an associated phone level transcription [6, 18].

3.4 Parameters Conversion of Speech Data

In order to build a set of HMMs for acoustic modeling, a set of speech data files and their associated transcriptions are required. The speech data file is converted into an appropriate parametric format (or the appropriate acoustic feature format) which consists of the required phone or word labels.

In this experiment, HMMs were trained using Mel Frequency Cepstral Coefficients (MFCCs) which are derived from FFT-based log spectra, consisting of the length of the parameterized static vector (+13), the delta coefficients (+13) and the acceleration coefficients (+13) which add up to 39 acoustic features. In order to code the data, a configuration file (config) is created which specifies all of the conversion parameters. The following is the setting used in this experiment.

```
# Coding parameters
SOURCEFORMAT = WAV          # Gives the format of the speech files
TARGETKIND = MFCC_0_D_A     # Identifier of the coefficients to use
# Unit = 0.1 micro-second:
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0       # = 25 ms = length of a time frame
TARGETRATE = 100000.0       # = 10 ms = frame periodicity
NUMCEPS = 12                # Number of MFCC coeffs (here from c1 to c12)
USEHAMMING = T              # Use of Hamming function for windowing frames
PREEMCOEF = 0.97            # Pre-emphasis coefficient
NUMCHANS = 26               # Number of filterbank channels
CEPLIFTER = 22              # Length of cepstral liftering
ENORMALISE = F
```

3.5 The Pronunciation Dictionary

The pronunciation dictionary, i.e., the lexical model, is one of the most important blocks in the development of large vocabulary speaker independent recognition systems [6]. In this paper, Tied-state triphones have been considered as the basic sub-word units that are extended from context independent phonemes. In the process of building a dictionary, the first step is to create a sorted list of the required words. The desired word list used to build a dictionary is extracted automatically from the text corpus using a script file that comes with HTK. Pronunciation dictionary that contains 1,448 unique words is prepared by taking words from a text corpus consisting of 250 sentences and phrases: 1248 words for training and 316 words for decoding (116 from training words and 200 words out of training words).

3.6 Language Model (LM)

Language model is another important requirement for any ASR system. LMs are used to constrain search in a decoder by limiting the number of possible words that need to be considered at any one point in the search. The consequence is faster execution and higher accuracy [19]. The LM used in this paper is a statistical language model, i.e., a backed-off bigram language model and the task grammar. Creation of a language model consists of computing the word uni-gram counts, which are then converted into a task vocabulary with word frequencies, generating the bi-grams from the training text based on this vocabulary, and finally converting the n-grams into a binary format language model and standard ARPA format [20]. It is developed using two of the CMU-Cambridge

statistical language modeling tools: the HLstats and HBuild. HLstats is used to gather and display statistical information for the label files and HBuild command is used to build the network.

4. Acoustic Model Training

Training the model means estimating the HMM's parameters, which are the mean, the variance, and the transition probabilities based on the utterance structure and the extracted parameters (features). The basic operation of the HTK training tools involves reading in a set of one or more HMM definitions, and estimating the parameters of these definitions using the speech data. Training consists of applying an embedded version of the Baum-Welch algorithm.

The first step in HMM training is to define a prototype model (with all means and variances initialized to 0 and 1 respectively). The topology of the model consists of the start and end states and three emitting states, using single Gaussian density functions. Then the initial parameter is estimated for the HMM model. It computes the global mean and variance and set all of the Gaussians in a given HMM to have the same mean and variance. Once the initial monophone model set is available, the next task is re-estimating the model parameters, i.e., estimating the transition probabilities of the Context-Independent HMMs.

So far, a monophone Acoustic Model is created that is not looking at the 'context' of the monophone [21]. The last step of model building is to transform the monophone HMMs into context-dependent triphone HMMs. With a triphone acoustic model, it is essentially looking for a monophone in the "context" of other monophones, i.e., the one immediately before and the one immediately after (if they exist - it may be the beginning or end of the word). There are 1883 unique triphones extracted from the training transcripts.

4.1 Recognizer Evaluation

After acoustic model training is completed, it is ready to evaluate the performance of the models. In this experiment, performance is assessed from five different speakers that do not participate in the training phase and from ten speakers that participated in the training phase. The accuracy of speech recognition system has always been measured by comparing the system's final output with a manual transcription performed by an individual.

The triphone-based systems have been developed in this experiment. For the tied-state triphone models, performance is evaluated for different Sp (duration between words) and also by changing only two of its parameters - the word insertion penalty (-p) and the grammar scale factors (-s) using task grammar and bigram as language model.

Tables 2 and 3 show the word recognition rates (%) for different Sp (duration between words) in milliseconds. The Pause between words can affect recognition accuracy whether the set time is too short or too long. It is better to speak at a normal pace like reading the news (avoid speaking one word at a time) since this improves accuracy for continuous speech. The pause setting of 200 milliseconds has better accuracy in this experiment.

Table 2: Triphone Tested Using Task Grammar

Sp duration in milliseconds	*WORD: % Correct*	*SENT: % Correct*	*Accuracy*
350	47.58	23.00	29.75
250	50.70	27.00	46.84
200	64.21	29.00	62.05
150	62.48	23.00	60.75

Table 3: Triphone Tested Using Bigram Language Model

Sp duration in milliseconds	*WORD: % Correct*	*SENT: % Correct*	*Accuracy*
350	50.48	0.00	9.47
250	53.68	0.00	25.67
200	60.75	0.00	29.44
150	57.72	0.00	27.27

The remaining test is done using the speech data with Sp duration of 200 milliseconds.

Table 4 shows the word recognition rates (%) when the recognition tool, HVite, is executed again on the triphone models by changing only two of its parameters: the word insertion penalty (-p) and the

grammar scale factors (-s). These parameters can have a significant effect on recognition performance. The –p value was incremented from 0.0 earlier to 20.0 and the –s value was taken up to 50.0.

Table 4: Triphone Tested Using Task Grammar

System	*WORD: % Correct*	*SENT: % Correct*	*Word Accuracy (%)*
M (-p=0.0, -s=5.0)	61.47	22.00	39.39
M (-p=0.0, -s=20.0)	63.35	29.00	61.47
M (-p=0.0, -s=25.0)	59.74	27.00	58.15
M (-p=0.0, -s=50.0)	36.65	13.00	36.51
M (-p=10.0, -s=20.0)	64.21	29.00	62.05
M (-p=20.0, -s=20.0)	64.21	29.00	61.62

All experiments above are performed without including QS command in the tree.hed file. *Tree.hed* contains the instructions regarding which contexts to examine for possible clustering.

QS - Question - is defined by the user and each QS command loads a single question and that question is defined by a set of contexts. Table 5 shows the word recognition rates (%) of the experiment with tree.hed containing Question that is generated manually to test the effects of *tying*.

Table 5: Triphone Tested with Tree.Hed Using Task Grammar

System	*WORD: % Correct*	*SENT: % Correct*	*Word Accuracy (%)*
Performance	27.71	11.00	24.68

The possible reason for this difference could be lack of complete and exhaustive information about the rules of Ge`ez. It would have been useful to determine the rules of Ge`ez that decide which consonants may cluster together in which order, and which are prohibited. AS *tree.hed* is generated manually to test the effects of *tying*, the rules specified in the script file *tree.hed* are by no means exhaustive in this study. This requires the involvement of scholars with strong linguistics background.

HVite is also executed on the triphone models by changing only the values of RO - the outlier threshold and TB that clusters one specific set of states. These values can have a significant effect on recognition performance as shown in Table 6. The RO value is changed from 100 earlier to 200 and the TB value was incremented from 350 to 750.

Table 6: Triphone Tested with tree.hed with Different RO and TB Values Using Task Grammar

System	*WORD: % Correct*	*SENT: % Correct*	*Word Accuracy (%)*
RO=100, TB=350	25.11	4.00	21.93
RO=200, TB=350	27.27	4.00	24.39
RO=200, TB=750	27.71	11.00	24.68

Finally, performance is done for untrained speakers. This test is performed using the speech data with Sp duration of 200 milliseconds as well as the -p value was 10 and the –s value was 20. Table 7 summarizes the recognition accuracy of speakers who do not participate in the training phase.

Table 7: Triphone Tested with Untrained Speakers

System	*WORD: % Correct*	*SENT: % Correct*	*Word Accuracy (%)*
Performance	62.19	24.00	57.72

5. Conclusion

This paper was devoted towards the development and evaluation of speech recognition system for continuous speech Ge'ez language of multiple speakers. In addition to researches done previously by other researchers which have great contributions on related local language, this experiment shows a promising result to design an applicable system that recognizes Ge`ez language speech with more enhancements.

However, this work is not exhaustive due to some factors such as limitation of freely available Ge'ez speech corpus to adequately train such models. This system is also based on the statistical approach of HMM. But it is possible to design better speech recognition system based on a recently introduced artificial neural network (ANN) technique and a hybrid system that combines ANN with other state-of-the-art techniques to improve performance

References

[1] Speech Recognition, -whitepaper- ASR,: http://www.docsoft.com/resources/Studies/ Whitepapers/whitepaper- ASR.pdf, Last accessed on June, 2017.

[2] Solomon Teferra Abate, Martha Yifiru Tachbelie, and Wolfgang Menzel, "Amharic Speech Recognition: Past, Present and Future", Proceedings of the 16th International Conference of Ethiopian Studies, Trondheim 2009.

[3] Kinfe Tadesse, "Sub-Word Based Amharic Word Recognition: An Experiment Using Hidden Markov Model (HMM)", 2002.

[4] Solomon Tefera Abate, Wolfgang Menzel, and Bahiru Tefla, "An Amharic Speech Corpus for Large Vocabulary Continuous Speech Re-Cognition", In Proc of 9th Europ. Conf. on Speech Communication and Technology, Interspeech, Lisbon, Portugal, 2005.

[5] Solomon Teferra Abate, Martha Yifiru Tachbelie, and Wolfgang Menzel, "Morpheme-Based Automatic Speech Recognition for Morphologically Rich Language – Amaharic", Department of Informatics, University of Hamburg Germany, https://pdfs. semanticscholar.org/.../3483737c4f0053dacf2fd 5e1b.

[6] Hafte Abera Miruts, "Hidden Markov Model Based Tigrigna Speech Recognition", Unpublished Masters Thesis, Addis Ababa University, 2009.

[7] Mesfin Birile, "Synthetic Speech Trained - Large Vocabulary Amharic Speech Recognition System", Unpublished Masters Thesis, Addis Ababa University, 2008.

[8] Michael Melese, Laurent Besacier, and Million Meshesha, "Amharic Speech Recognition for Speech Translation", Addis Ababa, 2016.

[9] Alula Tafera, "A Generalized Approach to Amharic Text-To-Speech (TTS) Synthesis System", Unpublished Masters Thesis, Addis Ababa University, 2010.

[10] Ge`ez Language, – Wikipedia the free encyclopedia, https://www.en.wikipedia.org/ wiki/Ge%27ez, Last accessed on May, 2017.

[11] Wolf Leslau, "Comparative Dictionary of Ge‟ez (Classical Ethiopic): Ge‟ez-English/English-Ge‟ez with an index of the Semitic roots," Wiesbaden Harrassowitz, 1991.

[12] Mulusew Asratie, "The Syntax of Non-verbal Predication in Amharic and Geez", LOT, The Netherlands, http://www.lotschool.nl, 2014.

[13] Pilar Quezzaire-Belle, "The Comparative Origin and Usage of the Ge'ez Writing System of Ethiopia", 2001, https://www.scribd.com/ document/50818987/paper-gabriella.

[14] www.persee.fr/doc/ethio_0066-2127_1957_num_2_1_1264.

[15] https://books.google.com/books?isbn =1575060191, 1997.

[16] ትንሣኤ ግእዝ, አፍሮ የሕትመት ስራ ድርጅት, 2012.

[17] Mohammad Abushariah, Raja Ainon, Roziati Zainuddin, Moustafa Elshafei, and Othman Khalifa, "Arabic Speaker-Independent Continuous Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus", Faculty of Computer Science and Information Technology, University of Malaya, Malaysia.

[18] Steve Young, Gunnar Evermann, and Dan Kershaw, "The HTK Book", Cambridge University, 2015.

[19] What-is-a-language-model, www.voxforge.org/ home/docs/faq/faq/, Last accessed on August, 2017.

[20] Speech recognition, – Wikipedia the free encyclopedia, https://en.wikipedia.org/wiki/ Speech_recognition, Last accessed on June, 2017.

[21] Create Acoustic Model Manually: Tutorial, 2017, www.voxforge.org, Last accessed on November, 2017.