

# Automatic Speech Recognition for an Under-Resourced Language - Amharic

Solomon Teferra Abate, Wolfgang Menzel

Department of Informatics, Natural Language Systems Group, University of Hamburg, Germany  
solomon\_teferra\_7@yahoo.com; menzel@informatik.uni-hamburg.de

## Abstract

In this paper we present the development of an Automatic Speech Recognition System (ASRS) for Amharic using limited available resources and the freely available speech toolkit (HTK). There are phonological, dialectal, orthographic and morphological features of Amharic that challenge the development of ASRSs. The problem of resource scarcity is also a hindrance to the research and development initiatives in the area of Amharic ASR. Dealing with these language and resource related problems, we have developed syllable- and triphone-based ASR for Amharic and achieved 90.43% and 91.31% word recognition accuracy, respectively, on the evaluation test set of 5k vocabulary.

**Index Terms:** speech recognition, under-resourced language, Amharic, syllable.

## 1. Introduction

Among thousands of languages in the world, only a small number possess the resources required for implementation of human language technologies, in general and speech technologies, in particular. Thus, these technologies are mostly developed for languages which have large resources or become interest of the economic or political influence. On the contrary, languages from developing countries or minorities have been less considered in the past. Amharic is one of these technologically disregarded languages. It is the official language of Ethiopia and a Semitic language that has the greatest number of speakers after Arabic in this family.

Only few attempts have been made in the area of Amharic natural language technologies, in general and Automatic Speech Recognition (ASR), in particular. All of the attempts, however, have suffered from shortage of the required resources [1].

Furthermore, the inflectional and derivational morphological feature of the language aggravates the problem of developing statistical models like statistical language model. In this paper we present the problems posed by the characteristics of Amharic and scarcity of resources in the development of ASR and our approach to deal with these problems.

## 2. Language Related Problems

### 2.1. Phonology

A set of thirty eight phones - seven vowels and thirty-one consonants - makes up the complete inventory of sounds for the Amharic language [2]. The consonants are generally classified as stops, fricatives, nasals, liquids, and semi-vowels [3]. Tables 1 and 2 show the classification of Amharic consonants and vowels<sup>1</sup>.

A labiovelar in any position followed by [ə] may become a plain velar followed by the labial round vowel [o] without any change in the meaning of the word [3]. Thus [qʷət'ə rə]

becomes [qot'ərə] - 'he counted'. Similarly, a labiovelar in any position followed by [I] usually becomes a plain velar followed by the labial rounded vowel [u]. For example, [qʷIrlsI] becomes [qurIsI] - 'breakfast'.

Although they are used only with the vowel [a], nearly all the other consonants can be pronounced with a slight rounding of lips. This requires proper handling in sub-word HMM modeling.

Table 1: Amharic Consonants

Manner of Art/n	Voicing	Place of Articulation				
		Labials	Dentals	Palatals	Velars	Glottals
Stops	Vs	[p]	[t]	[tʃ]	[k]	[ʔ]
	Vd	[b]	[d]	[dʒ]	[g]	
	Glott	[pʰ]	[tʰ]	[tʃʰ]	[q]	
	Rd				[kʷ][gʷ] [qʷ]	
Fricatives	Vs	[f]	[s]	[ʃ]		[h]
	Vd		[z]	[ʒ]		
	Glott		[sʰ]			
	Rd					[hʷ]
Nasals	Vd	[m]	[n]	[ɲ]		
Liq	Vd		[l] [r]			
Sv	Vd	[w]			[j]	

Key: Glott = Glottalized; Vs = Voiceless; Vd = Voiced; Rd = Rounded; Liq = Liquids; Sv = Semi-Vowels.

Table 2: Amharic Vowels

Positions	front	center	back
high	[i]	ɨ [I] is used	[u]
mid	[e]	[ə]	[o]
low		[a]	

### 2.2. Dialects

Amharic has five dialectal variations spoken in five different Amharic speaking regions: Addis Ababa, Gojjam, Gonder, Wollo, and Menz [5]. The speech of Addis Ababa has emerged as the standard dialect and has wide currency across all Amharic-speaking communities [6]. The dialects vary at phonological and morphological levels, which need to be handled in the development of an ASRS that can be used by all the dialect speakers of the language.

### 2.3. Writing System

Amharic writing system is used across all Amharic dialects. Getachew [7] states that the Amharic writing system is phonetic. Leslau [3] also noted that in general, no real problems exist in Amharic orthography, as there is more or less, a one-to-one correspondence between the sounds and the graphic symbols, except for the geminated consonants

<sup>1</sup>International Phonetic Association's (IPA) standard has been used for representation.

and redundant symbols. Many ([8], [5], [9]) have claimed the Amharic orthography as a syllabary in that a grapheme represents a concatenated Consonant and Vowel (CV) syllable.

The phonetic and syllabic features of the Amharic writing system present advantages for the development of ASR. They ease the development of pronunciation dictionaries by using grapheme-based approach.

However, Amharic orthography does not indicate gemination that range from a slight lengthening to much more than doubling of consonants. In Amharic the [I] vowel and the glottal stop consonant [ʔ] are not realized in all words. But the orthography does not show the irregularities in the pronunciations of these phones.

Although the Amharic orthography (as represented in the Amharic character set, also called [fidAll]) consists of 276 distinct symbols, there are redundant symbols that represent the same sounds. Four graphemes  $\upsilon$ ,  $\mathfrak{h}$ ,  $\mathfrak{r}$  and  $\mathfrak{n}$  representing the [h] (as in he) sound, two graphemes  $\mathfrak{w}$  and  $\mathfrak{h}$  represent the [s] (as in speech) sound, two graphemes  $\mathfrak{h}$  and  $\mathfrak{h}$  depict the [a] (as in an) sound and two graphemes  $\mathfrak{x}$  and  $\mathfrak{h}$  depicting the [s'] (which has no equivalent sound in English) sound. Each of these graphemes has seven forms (called orders as given in Table 3). All the graphemes with red colour also represent the same sound with  $\upsilon$ . When these redundant graphemes with their orders are eliminated, only 233 distinct CV syllable characters remain.

Table 3: Redundant Graphemes of the [h] sound

	$\mathfrak{e}$	$\mathfrak{u}$	$\mathfrak{i}$	$\mathfrak{a}$	$\mathfrak{e}$	$\mathfrak{l}$	$\mathfrak{o}$
h	$\mathfrak{u}$	$\mathfrak{v}$	$\mathfrak{z}$	$\mathfrak{y}$	$\mathfrak{z}$	$\mathfrak{v}$	$\mathfrak{v}$
h	$\mathfrak{h}$	$\mathfrak{h}$	$\mathfrak{h}$	$\mathfrak{h}$	$\mathfrak{h}$	$\mathfrak{h}$	$\mathfrak{h}$
h	$\mathfrak{r}$	$\mathfrak{r}$	$\mathfrak{r}$	$\mathfrak{r}$	$\mathfrak{r}$	$\mathfrak{r}$	$\mathfrak{r}$
h	$\mathfrak{n}$	$\mathfrak{n}$	$\mathfrak{n}$	$\mathfrak{n}$	$\mathfrak{n}$	$\mathfrak{n}$	$\mathfrak{n}$

## 2.4.Morphology

Amharic is one of the morphologically rich languages. It has inflectional and derivational morphology. These morphological features of Amharic, considerably increase the number of word forms which can be derived from the available roots.

The morphological property of Amharic keeps the frequency of word forms very low. In a text of about 114,267 words, we have observed that 52% of the words appear only once while only 0.05% appear more than 1000 times.

To have a comparative view of this property of the language, we compared an English text with 13,126 different words with an Amharic one of 24,284 different words and show their word frequencies in Table 4. The English text is taken from WSJCAM0.

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

Table 4: Frequency of Words in Amharic and English Texts

Frequency	Amharic		English	
	# of words	Percent	# of words	Percent
$\geq 1000$	1	0.004	52	0.40
100-999	48	0.2	550	4.2
10-99	1,456	6.0	3,362	25.6
2-9	6,233	25.7	8,227	62.7
1	16,547	68.1	932	7.1

## 3.Resource Related Problems

### 3.1.Speech Corpus

The most fundamental resource for any speech recognition research and development is speech corpus. However, out of the different types of speech corpora, we have only a read speech corpus for Amharic. In this corpus, the Addis Ababa dialect is covered better than the other dialects [10].

The corpus is only a medium size speech corpus of 20 hours of speech [10]. Compared to other speech corpora that contain hundreds of hours of speech data for training, the Amharic speech corpus illustrates how badly Amharic is under-resourced. For example, the British National Corpus contains 1,500 hours of speech, CGN (Spoken Dutch Corpus) has total of 800 hours of speech and 104 hours of read-speech, CSJ (Corpus of Spontaneous Japanese) consist of 650 hours of speech, and BREF-120 is a large corpus of French read speech with 100 hours of speech.

### 3.2.Pronunciation Dictionary

Although the pronunciation dictionary is a key component of any ASR system, no such dictionary is available for Amharic. To prepare one, the following options can be considered [11]:

- knowledge-based approaches used to build a phonetizer;
- automatic approaches using a phone recognizer;
- grapheme-based approaches.

Since there is no human and financial resource to apply the first and the second methods, our pronunciation dictionary is developed using the third method [10]. In the transcription of the graphemes of the word, it has been assumed that a grapheme represents a CV syllable sound. The existence of any other structure of syllables in the language is disregarded. This method, therefore, does not handle the gemination of consonants and the irregular realization of the sixth order vowel and the glottal stop consonant that are not indicated in the writing system of the language.

### 3.3.Language Model

Training a statistical language model requires text data that consist of millions of words [12]. However, this requires a bulk of text pre-processing which includes manual correction of grammar and spelling because there is no Amharic grammar or spelling checker. The other problem in the development of an Amharic language model comes from the morphological richness of the language. For such a language researchers, like [13], suggest the development of a sub-word language model, which, however, requires a well performing sub-word parser.

## 4. Development of the ASRSs

As it has been indicated in the previous section, there is no properly developed pronunciation dictionary and language model and we have only a medium size speech corpus for Amharic. Therefore, we have used different approaches to deal with the above-mentioned language and resource related problems in the course of developing an ASRS for Amharic. This section presents our approaches in the development of pronunciation dictionaries, language and acoustic models. Since we have used syllable and triphone as recognition units, our pronunciation dictionaries and acoustic models are developed using Amharic CV syllables and phones.

### 4.1. Pronunciation Dictionaries

The pronunciation dictionaries that we used for our experiments have been developed by a simple transcription of words in terms of concatenated CV syllables and separated phones for the syllable- and phone-based dictionaries, respectively.

To investigate the influence of the irregularities in the pronunciation of the glottal stop consonant, we have developed two phone-based dictionaries: one assumes realization of this consonant whenever it is written and the other ignores its realization. Our triphone-based recognizers benefited from the use of the latter dictionary.

We have also developed two other pronunciation dictionaries to deal with the difference between the rounded (labiovelars) and unrounded consonants. One uses two different symbols for rounded and unrounded consonant, for example, the consonant [h] is represented by the symbols hue when it is rounded and h when it is not. In the second dictionary, the rounding feature is attached to the next vowel instead of the consonant, for instance, the syllable [ha] is represented as h aue when its consonant is rounded and h a otherwise. A triphone-based recognizer that uses the latter version of our pronunciation dictionary achieved a better word recognition accuracy.

### 4.2. Language Models

Since there is a clean training text consisting of less than 750,000 words, we have limited our language models to bi-grams instead of opting for tri-gram or more. Even this text is not enough to properly train a good bi-gram language model. To minimize this shortage of training text, we have included sentences from the other test sets for developing a language model for a test set. For example, when we develop the language model for the 5,000 development test set, we included sentences from the other test sets during training. The sentences in this test set could also be used to train the language model for the other test sets, thus increasing the training text size to approximately 75,000 sentences and 900,000 words.

The language models developed in this experiment are closed vocabulary model which are limited to the given test vocabulary during development. We have used the HTK tool to develop the language models and measure their perplexities.

### 4.3. Acoustic Models

#### 4.3.1. Triphone-based Model

For the triphone-based systems an HMM with a 3-state left-to-right topology was used. Since there is no speech data that is segmented at phone level, we used monophone models

that have been initialized with the flat start procedure to develop our intra-word triphone models.

The shortage of training speech data has partly been compensated by tying the components of the triphones and using diagonal instead of full covariance matrices. We used decision trees to cluster states and then tie each cluster. The use of decision trees is based on asking questions about the left and right contexts of each triphone. Tying enabled many triphone models to share training speech data, thereby reducing the number of triphone models from 5,092 logical models to 4,099 physical models.

To find out if the data is sufficient to train models with skips, we have developed different recognizers using HMMs with skips and others without skips. The best versions achieved 90.94% (with 12 Gaussian mixtures) and 90.46% (with 20 Gaussian mixtures) word recognition accuracy. This indicates that the data was sufficient to train the transition parameters for the skips, but did not allow us to increase the number of Gaussian mixtures to more than 12.

To deal with the irregular realization of the vowel [I] and the glottal stop consonant [ʔ], we also conducted an experiment using a jump from the first emitting state to the final non-emitting state for these phones. The trained transition probabilities of the models show that the data indeed favour this instead of moving to the next emitting state. Additionally, we have assumed that the problem of gemination may be compensated by the looping state transitions of the HMMs.

#### 4.3.2. Syllable-based Model

To test the utility of the speech data that is segmented at syllable-level, we have initialized syllable-based HMMs with both the bootstrapping and the flat start method and trained them in the same way. The HMMs that have been initialized with the flat start method performed better (40% word recognition accuracy) on development test set of 5,000 words. With the understanding that the segmentation has some problems, we have conducted all consecutive experiments on the models that have been initialized with the flat start method.

To develop a good HMM model for Amharic CV syllables as recognition units, one needs to conduct experiments to select the optimal model topology. Designing an HMM topology consists of choosing an appropriate number of states, the allowed initial states and the allowed transitions. That has to be done with proper consideration of the size of the unit of recognition and the amount of the training speech data. This is due to the fact that as the size of the recognition unit increases and the size of the model (in terms of the number of states and number of transitions) grows, the model requires more training data.

We, therefore, carried out a series of experiments using a left-to-right topology with and without jumps and skips, with a different number of emitting states (3, 5, 6, 7, 8, 9, 10 and 11) and different number of Gaussian mixtures (from 2 to 98) to get the optimal topology that can address language and resource related problems. By jump, in the case of the syllable-based modeling, we mean skips from the first non-emitting state to the middle state and/or from the middle state to the final non-emitting state.

To deal with resource related problems, we have limited the number of states to be skipped to one because the amount of training speech that we have is too limited to train the additional transition probabilities for skipping two or more states. Diagonal covariance matrices instead of full covariance matrices have been used to perform reliable re-estimation of the components of the model with our limited training data. Unlike the triphone model, tying has not yet

been applied on the syllable models to deal with the problem of data shortage.

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

To solve the problems of the irregularities in the realization of the sixth order vowel [I] and the glottal stop consonant [ʔ], HMM topology with jumps has been used. Unlike the triphone models, a jump from the middle state to the final non-emitting state for all of the CV syllables with the sixth order vowel, and a jump from the first emitting state to the middle state for all of the CV syllables with the glottal stop consonant have been used. It has been observed from the transition probabilities of the trained models that they favor such a jump. For example, the models of the glottal stop consonant with the sixth order vowel tend to jump from the first non-emitting state to the 4<sup>th</sup> state (with probability of 0.72) instead of moving to the first emitting state (that has transition probability of 0.38). They also tend to jump from the 4<sup>th</sup> state to the final non-emitting state (with probability of 0.28) instead of transitioning to the next state (which has transition probability of only 0.10).

#### 4.4. Recognition Results

We present recognition results of only those recognizers which have best performance on the development test set from all our syllable-and triphone-based recognizers. We have also given only the best results of our evaluation on the evaluation test sets. Therefore, Table 5<sup>2</sup> presents word recognition accuracy of two recognizers that have best performance on the 5k development and evaluation test sets. Both recognizers used 12 Gaussian mixtures per state.

Table 5: Word Recognition Accuracy

Test Sets	HMMs	Models	
		AM+LM	AM+LM+SA
Development	Syllable	88.99	89.80
	Triphone	90.94	90.75
Evaluation	Syllable		<b>90.43</b>
	Triphone	<b>91.31</b>	

The models with 3 emitting states, 12 Gaussian mixtures and with skips achieved the best word recognition accuracy (91.31%) among the triphone-based recognizers that we have developed. Among the syllable-based models those with five emitting states, with 12 Gaussian mixtures and without skips and jumps turned best with a word recognition accuracy of 90.43%. From the point of view of their word recognition accuracy, the triphone model performed better than the syllable model. But we have to take note that the syllable model is not tied at all and may gain more from tying than the triphone model.

<sup>2</sup>In the table, AM stands for Acoustic Model, LM for Language Model and SA for Speaker Adaptation.

## 5. Conclusions and Research Directions

Despite the problems that are presented in sections 2 and 3, the performance of our recognition systems is encouraging. We still see that there are promising lines of performance improvement without considering a labor intensive extension of the speech and text corpora. To mention a few: tying the parameters of syllable-based models; adaptation of speakers from the different dialects of the language; development of a properly trained language model by applying different smoothing techniques, using sub-word modeling units etc.

We can learn that proper utilization of the available techniques in the ASR technology minimizes the expense of collecting and annotating speech corpora for under-resourced languages. The use of grapheme-based approach in the development of pronunciation dictionaries and acoustic models is applicable to all the Ethiopian languages that use the same writing system.

## 6. References

- [1] Atelach Alemu, Lars Asker and Mesfin Getachew, "Natural Language Processing for Amharic: Overview and Suggestions for a Way Forward" TALN 2003, Batz-sur-Mer, 11-14, 2003.
- [2] Baye Yimam and TEAM 503 students, "ፊደል አንጻራዊ" Ethiopian Journal of Languages and Literature 7(1997):1-32, 1997.
- [3] Leslau, W., "Introductory Grammar of Amharic", Wiesbaden: Harrassowitz, 2000.
- [4] Cohen, M., N.V.Yushmanov and E. Ullendorff, "An Amharic Chrestomathy", London: Oxford University Press, 1965.
- [5] Cowley, Roger, Marvin L. Bender and Charles A. Fergusone, "The Amharic Language-Description" In Language in Ethiopia, London: Oxford University press, 1976.
- [6] Hayward, Katrina and Richard J. Hayward "Amharic", In Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge: the University Press, 1999.
- [7] Getachew Haile, "The Problems of the Amharic Writing System". A paper presented in advance for the interdisciplinary seminar of the Faculty of Arts and Education. HSIU, 1967.
- [8] Bender, L.M. and Ferguson C., "The Ethiopian Writing System". In Language in Ethiopia, London: Oxford University press 1976.
- [9] Baye Yimam "የአማርኛ ሰዋሰው", Addis Ababa. ት.መ.ጥ.ጥ.ድ. 1986.
- [10] Solomon Teferra Abate, Wolfgang Menzel and Bairu Tafla, "An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition", In: INTERSPEECH 2005, 2005.
- [11] Besacier, L. V-B. Le, C. Boitet, V. Berment. "ASR and Translation for Under-resourced Languages". Proceedings IEEE ICASSP 2006. Toulouse, France. May 2006.
- [12] Jelinek, F, "Self-organized language modeling for speech recognition" In A. Waibel and K.F. Lee, editors, Readings in Speech Recognition: 450-506. Morgan Kaufmann, 1990.
- [13] Kirchhoff, Katrin et al., Novel Speech Recognition Models for Arabic. [http://ssli.ee.washington.edu/people/katrin/arabic\\_resou rces.html](http://ssli.ee.washington.edu/people/katrin/arabic_resou rces.html), 2002