An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition

Solomon Teferra Abate, Wolfgang Menzel, Bairu Tafila,

Fachbereich Informatik, Universität Hamburg
email : {solomon, menzel}@nats.informatik.uni-hamburg.de

Abstract

Amharic is the official language of Ethiopia. It belongs to the Semitic language family and is characterized by a quite homogeneous phonology distinguishing between 234 distinct Consonant-Vowel (CV) syllables.

Since there is no Amharic speech corpus of any kind, we developed a read-speech corpus using a phonetically rich and balanced text database. To prepare the text database, we used the archive of EthioZena website which consists of selected articles from well known newspapers and magazines published in Amharic. The archive was cleaned semi-automatically.

Like other standard speech corpora, such as WSJCAM0, the Amharic speech corpus contains training set, speaker adaptation set, test sets (development and evaluation test sets each with 5000 and 20000 vocabulary size). The speech has been recorded in Ethiopia in an office environment and segmented semi-automatically. The corpus is now used for experiments with a syllable- and phone-based LVCSR for Amharic.

1. The Amharic Language

Amharic is the official language of Ethiopia. It is a Semitic language family that has the largest number of speakers after Arabic [1]. It is spoken, as per the 1998 census, by 17.4 million people as a mother tongue and 5.1 million people as a second language [2]. Amharic has five dialectical variations spoken in different Amharic regions: Addis Ababa, Gojjam, Gonder, Wollo, and Menz [3]. The dialect of Addis Ababa has emerged as the standard dialect and has wide currency across all Amharic-speaking communities [1].

Like other languages, Amharic has its own characterizing properties. For example, Amharic has a set of speech sounds that are not found in other languages, for example English. These are the glottalized plosives ጵ, ጥ, ጽ, ጭ, and ቅ which have a sharp click-like characters [4]. A review of the Amharic phonetics, morphology and writing system is given in view of speech corpus development.

1.1. Amharic Phonetics

A set of 38 phones, seven vowels and thirty-one consonants, makes up the complete inventory of sounds for the Amharic language [5]. We give a brief overview of each of these major categories of Amharic phones.

1.1.1. Consonants

Amharic consonants are generally classified as stops, fricatives, nasals, liquids, and semi-vowels [4]. Table 1 shows the phonetic representation of the consonants of Amharic as to their manner of articulation, voicing, and place of articulation[1].

Manner of Art/n	Voicing	Place of Articulation				
		Lab	Dent	Pal	Vel	Glo
Stops	Voiceless	ፕ[p]	ት[t]	ች[tʃ]	ክ[k]	አ[?]
	Voiced	ብ[b]	ድ[d]	ጅ[dʒ]	ግ[g]	
	Glottalized	ጵ[p‘]	ጥ[t‘]	ጭ[tʃ‘]	ቅ[q]	
	Rounded				ኵ [k^w] ጕ[g^w] ቍ[q^w]	
Fricatives	Voiceless	ፍ[f]	ስ[s]	ሽ[ʃ]		ህ[h]
	Voiced		ዝ[z]	ዥ[ʒ]		
	Glottalized		ጽ[s‘]			
	Rounded					ኍ[h^w]
Nasals	Voiced	ም[m]	ን[n]	ኝ[ɲ]		
Liquids	Voiced		ል[l] ር[r]			
Semi vowels	Voiced	ው[w]			ይ[j]	

Table 1: *Categories of Amharic Consonants. Keys: Lab is Labial; Den is Dental; Pal is Palatal; Vel is Velar; Glo is Glottal*

1.1.2. Vowels

The vowels (ኧ, ኡ, ኢ, ኣ, ኤ, እ, and ኦ) are categorized as rounded (ኡ and ኦ) and unrounded (ኣ, ኤ, እ, ኢ and ኧ). Another categorization according to the place of articulation is given in table 2.

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

Table 2: *Categories of Amharic Vowels.*

1.2. Amharic Morphology

Amharic has a fairly rich morphology marking, for example, pronominalized verbal arguments at the verb:

1. The subject is marked on the verb using subject suffix pronouns, as in ሰበርኩ - ’I broke’;

[1]Here and in table 2 all the symbols in the square bracket are IPA representations of Amharic phones

2. The direct object is optionally marked on the verb, as in ሰበረኝ - 'he broke me';
3. Some prepositional phrase complements are optionally marked on the verb as in እስኪሰብረኝ - 'until he broke me';
4. Functional elements like negation marks, conjunctions and some auxiliary verbs are also bound morphemes and are attached to the verb. For example, in አልሰበርም 'I will not be broken', the negation is marked by አል.

All arguments of a verb are optional and may only be indicated by suffix pronouns, that is, a verb may stand alone as a sentence.

The definite article in Amharic is also a bound morpheme and is attached to a noun or to the first inflected element in a noun phrase.

This high morphological variety of the language increased the size of the pronunciation dictionary considerably. The verb ጀመረ 'to begin', for example, has 103 different forms in our training dictionary of 28666 words and 167 different forms in another dictionary of 51,489 words . Some of its forms are given below.

ጀመራችሁ	ጀ መ ራ ች ሁ	'You started'
ጀመረ	ጀ መ ረ	'He started'
ጀመረች	ጀ መ ረ ች	'She started'
ጀመርኩ	ጀ መ ር ኩ	'I started'
ጀመርን	ጀ መ ር ን	'We started'
ጀመሩ	ጀ መ ሩ	'They started'
የጀመሩትን	የ ጀ መ ሩ ት ን	'The one that they started'

All the verbs have the potential for such a variation. In addition nouns and pronouns also carry different elements of a sentences such as conjunctions, articles, possession marks and prepositions.

2. The Amharic writing system

Getachew [6] stated that the Amharic writing system is a phonetic one in that it allows any one to write Amharic texts if s/he can speak Amharic and has knowledge of the Amharic alphabet. In support of the above point, [7] noted that no real problems exist in Amharic orthography, as there is more or less, a one-to-one correspondence between the sounds and the graphemes, except redundant graphemes.

For a relatively long period of time many authors [8]; [5] have claimed that Amharic orthography is syllabic. Only recently Taddesse [9] and Baye [10] have argued that it is possible to represent Amharic speech using either isolated phoneme symbols or concatenated CV syllabic symbols. In the concatenated feature, which is the common and known to the most population, each orthographic symbol represents a consonant and a vowel. This is the basis for the claim that the Amharic writing system is syllabic.

The Amharic orthography as it is represented in the Amharic character set, called ፊደል, consists of 276 distinct symbols. These symbols are classified into four groups. In the first category (33*7=231) there are thirty-three core orthographic symbols, each of which has seven different shapes, usually known as orders, to represent the seven vowels. Each consonant and the seven vowels in combination represent CV syllables. The second category (4*5=20) consists of four labio-velar symbols, which have five orders. The eighteen labialized consonants, which have only one order, are the third category. The fourth category is the grapheme ቭ with its 7 orders.

In Amharic there are four graphemes (ህ ሕ ኅ ኽ) representing the /h/ sound and two graphemes (እ ዕ) that represent the /?/ sound. Their 1st and 4th order graphemes, except the grapheme (ኽ), also represent the same sound. There is a special grapheme ኧ that stands for the first order of the /?/ sound, two graphemes (ሥ ስ) represent the /s/ sound and two graphemes (ጽ ፅ) represent the /s'/ sound. Since in developing a speech recognition system we focus on modeling distinct sounds, only one of the redundant graphemes needs to be used to represent a sound. This reduces the number of graphemes necessary to represent distinct CV syllables to 234. As the first order of the /?/ sound appears only in the word ኧረ 'why!', we excluded it from the corpus. Finally we are left with 233 distinct CV syllable sounds.

2.1. Motivation of Amharic Speech Corpus Preparation

Speech recognition research in the major languages like English, German and Chinese has been conducted since the 1950s. But only recently, a few experimental attempts [11], [12], [13], [14] have been made for Amharic. These activities suffered from the lack of an Amharic speech corpus producing a demand for collecting a speech corpus suited for the development of an Amharic speech recognizer. This was also our problem in exploring the possibilities of developing large vocabulary, continuous speech recognition (LVCSR) for Amharic. We needed, therefore, to develop a speech corpus.

3. Preparation of Amharic Speech Corpus

Due to financial and time constraints this project was limited to the preparation of a read-speech corpus.

It has been designed according to best practice guidelines established for other languages. Standard speech corpora, such as the Wall Street Journal speech corpus [15], consist of training set, speaker adaptation set and test sets (development and evaluation test sets each with 5000 and 20000 vocabulary size). To make it comparable with commonly used standard corpora, the Amharic corpus has been made to contain the same components. A brief description of the preparation of the text database for these components is given in subsection 3.1.

3.1. Text Corpus Preparation

In contrast to other languages like English, there are no easily available electronic text sources for Amharic. Fortunately, the archive of EthioZena website was made available to us in a special encoding called SERA in HTML. From this archive more than 100,000 sentences have been acquired. To fit with other processing components the text was converted from HTML to plain text and from SERA to ethiop encoding [16]. To clean the acquired text:

- Spelling and grammar errors have been corrected;
- Abbreviations have been expanded;
- Foreign words have been eliminated;
- Numbers have been textually transcribed; and
- Concatenated words have been separated;

After having a clean text database:

- A total of 72,000 sentences with a maximum of 20 words in length have been chosen from the clean text. The collection of these sentences is used as a text database;

- From this text database, 53 sentences have been selected to create a phonetically balanced text database for the speaker adaptation set;
- From the remaining data in the text database, 10,000 sentences have been selected to create a phonetically balanced database for the training set;
- From the remaining data in the text database 850 and 3000 sentences have been selected randomly for the 5,000 and 20,000 vocabulary evaluation test sets, respectively. Similarly 1000 and 4000 sentences have been selected for the 5,000 and 20,000 vocabulary development test sets, respectively;

The selection of phonetically balanced sentences was done automatically in two steps. First the so-called important sentences [17] were selected. These sentences were intended to include all the syllables that are in the database. Secondly sentences that enrich the phonetic balance of the recording sets were selected based on an add-on procedure. Based on the distribution of syllables in the current selection, a score that estimates the utility of a sentence for achieving the desired target distribution is computed. The sentence with the highest utility is chosen and deleted from the database [17].

In the selected sentences, 19 syllables were missing and a few others have too low a frequency to be used in a statistical method like Hidden Markov modeling. We solved this problem by collecting Amharic words that contain missing and rare syllables in consultation with language experts choosing only words which are active in modern Amharic. Sentences are constructed using these words according to the grammar of the language. The frequency distribution of the syllables in the original and the modified speech corpus is given in figure 1 and table 3.

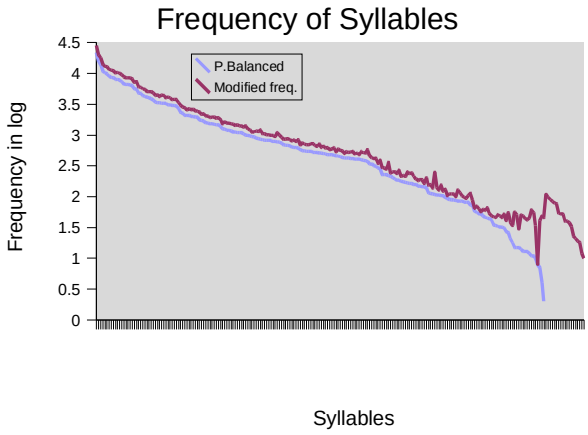

Figure 1: Frequency of Syllables

3.2. Recording of the speech

Having the text corpus available, the next important step in read-speech corpus preparation is recording the speech. We have carried out a supervised on-site recording [18]. The recording has been done in Ethiopia in an office environment using a laptop computer and a headset close speaking, noise canceling microphone. The age distribution of all the speakers is shown in Table 4.

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

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

Table 3: *Improvement in syllable frequency. Keys: Syll stands for Syllables; PB.Freq stands for Phonetically balanced frequency and Mo.Freq stands for the modified frequency*

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

Table 4: *Age distribution of the Readers.*

The training set consists of a total of 10850 different sentences. It includes 450 sentences that are necessary to consider missing syllables and enrich the frequency of the rare ones. The training set was read by 80 speakers of the Addis Ababa dialect, 70 of them read 100 sentences each and 10 of them read 145 sentences each. We had also recorded speech of 20 speakers of the other four dialects, who read 120 sentences (100 from the phonetically balanced training text database and 20 sentences from the constructed sentences) each, for the training set.

Test and speaker adaptation sets were read by 20 other speakers of the Addis Ababa dialect and 4 speakers of the other four dialects. For the 5000 vocabulary (development and evaluation) and the 20000 vocabulary (development and evaluation) test sets, 18 and 20 different sentences have been selected, respectively for each speaker.

For the recording purpose one sentence at a time was displayed to the speaker to read. The whole recording was done

[3]Due to time constraints for recording in the respective regions and difficulties to find enough dialect speakers in Addis Ababa, we have not been able to keep the age balance for the dialect speakers

in the presence of the researcher who controlled the recording. The control included: running the recording script, starting the recording session, breaking the recording, playing back the recorded speech, re-recording the sentence (if required) and moving to the next sentence. Every speaker was explained the purpose of the project and instructed to start the recording when s/he was ready to read. After the entire session for a reader was finished, all the utterances were listened to by the reader and the researcher for corrections. Furthermore, as only 4-5 speakers have been scheduled to read per day, the researcher was able to listen again to the recorded speech at the end of each day to control its quality.

The corpus is annotated manually and segmented semi-automatically at word and syllabel level, respectively.

The speech corpus contains 20 hours of training speech collected from 100 speakers who read a total of 10850 sentences. The Total number of words in the training speech is 28666 that covers all the 233 Amharic CV syllables.

4. Availability

So far the corpus is used primarily for our research of developing automatic speech recognizer for Amharic. Only few students at Addis Ababa university, faculty of Informatics used a portion of the corpus for their work of master thesis projects. As our research is at its last phase, we have an intention of making the corpus available for researchers and developers through a third party who may show an interest to do so.

5. Acknowledgments

We are grateful to Daniel Yacob who made the archive of EthioZena available to us. He also provided his SERA to Ethiop (namely g2) font conversion tool. Without his support the project would have been really expensive and time consuming.

Most of all we would like to thank DAAD (Deutscher Akademischer Austauschdienst) for generously providing the required funds for speech data collection as well as supporting one of the authors through a PhD grant.

6. References

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

[2] Ethnologue. 2004. Languages of the World, 14th Edition. $http : //www.ethnologue.com/14/show_language.asp?code = AMH$

[3] Cowley, Roger, et al. 1976. The Amharic Language, in Bender, M. et al (eds.), Languages in Ethiopia. London: Oxford University Press.

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

[5] Baye Yimam. 1986. "የአማርኛ ሰዋሰው". Addis Ababa ት. ሙ. ማ. ማ. ድ.

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

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

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

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

[10] Baye Yimam and TEAM 503 students. 1997. "የአማርኛ ሰዋሰው" Ethiopian Journal of Languages and Literature 7(1997): 1-32.

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

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

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

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

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

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

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

[18] Schiel, Florian and Christoph Draxler. 2003. Production and Validation of Speech Corpora: Bavarian Archive for Speech Signals. printed by Books on Demands GmbH, Norderstedt.