DEVELOPMENT OF AUTOMATIC AMHARIC SPEECH RECOGNIZER

Yitagessu Birhanu Gebremedhin, Rüdiger Hoffmann

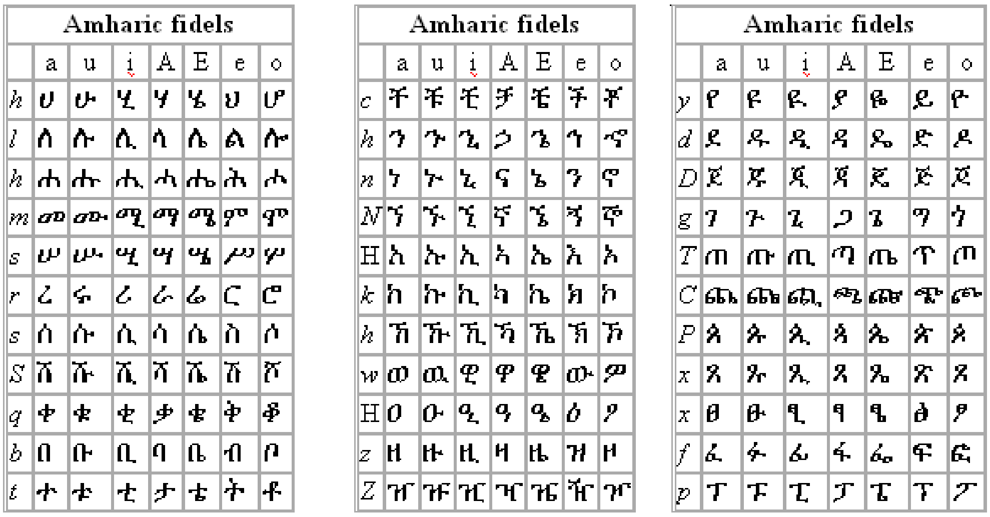
Dresden University of Technology, Institute of Acoustic & Speech Communication
Yitagessu_Birhanu.Gebremedhin@mailbox.tu-dresden.de , *Ruediger.Hoffmann@tu-dresden.de*

Abstract: Amharic is one of the least researched languages in the world. Particularly, speech and language technologies for this language are almost non-existent. Off the rack speech corpuses, lexical models and language models are not available and this makes the task of building automatic Amharic speech recognizers very challenging. We present initial results in the development of Amharic speech recognizer. The most important components of the speech recognizer; namely the speech corpus, the lexical model and the language model are developed from scratch. The Amharic speech corpus was collected from people of different age range and gender in such a way that it has all the syllables in an approximately reasonable proportion. A lexical model consisting of hundreds of thousands of words and a Finite-State-Automata based language model are also prepared. The speech recognizer is being developed using the UASR (Unified Approach to Speech Synthesis and Recognition) toolkit of TU Dresden and when it is ready we will integrate it with other modules for further research work, particularly in the development of Amharic speech to Ethiopian Sign Language (ESL) converter.

Keywords: UASR, dlabpro, language mode, lexical model, acoustic model, FSM, GRM, Audacity, wavesurfer

1. The Amharic Language

Amharic is the official language of Ethiopia and the second widely spoken Semitic language in the world (Hayward and Richard 1999). It is written from left to right using a scripting system known as Ge'ez. This scripting system is unique in that the consonants and vowels in its orthography do not exist independent of each other. The alphabet consists of 7 vowels and 33 consonants; each of the latter has 7 shapes depending on the vowel with which it is combined.



Amharic fidels							
	a	u	i	A	E	e	o
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
h	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
s	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
S	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ

Amharic fidels							
	a	u	i	A	E	e	o
c	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ
h	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
N	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ
H	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
h	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ
w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
H	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
Z	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ

Amharic fidels							
	a	u	i	A	E	e	o
y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
D	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
T	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
C	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ
P	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
x	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
x	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

Fig. 1: Amharic syllables

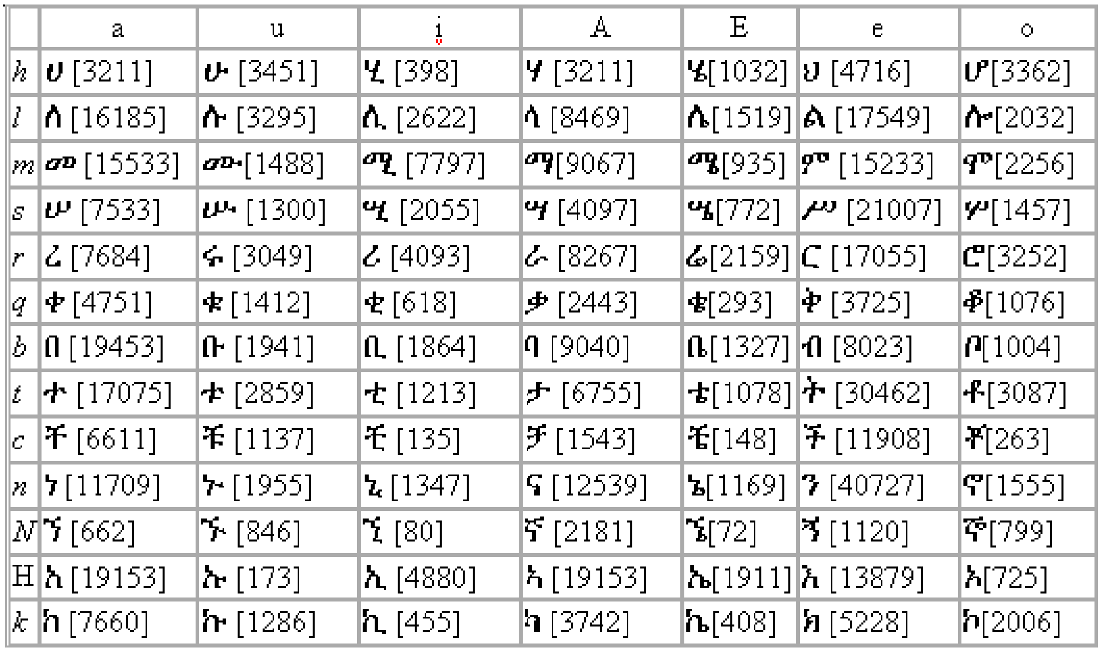
2. Related Works

Previous attempts to build automatic Amharic speech recognizers are very limited. Solomon [1] built both speaker dependent and independent, isolated consonant-vowel (CV) syllable- based recognition systems for Amharic. His work was extended to an isolated word recognizer by Kinfe [2]. He collected a 170 words vocabulary from 20 speakers and considered a subset of the Amharic syllables, concentrating on the combination of 20 consonants with the seven vowels. Both his training and test sets consist of 50 discrete words. He reported an isolated word recognition accuracy of 83.1% and 78.0%, for speaker dependent; phoneme and tied-state triphone models, respectively. Martha [3] developed a small vocabulary, isolated word recognizer for command and control interface to Microsoft Word. Zegaye [4] extended Martha’s work and developed a speaker independent continuous Amharic speech recognizer. He reported 76.2% word accuracy and 26.1% sentence level accuracy. Finally, the most impressive research work of all attempts is Solomon’s [5] large vocabulary, speaker independent, continuous speech recognition system. He reported a word recognition accuracy of 90.43% and 91.31% for syllable based and tri-phone based systems, respectively.

3. The speech corpus

We prepared a speech corpus of about 30 Hours duration. The training set is collected from 133 speakers of different age range and gender and the test set is collected from 15 people. We attempted to make the training set as phonetically balanced as possible so that the corpus has the syllables in a fairly equal proportion.

Sixty of the speakers in the training set are female and the rest are male. The minimum number of sentences read by a single female speaker is 83 and the maximum is 149. In a similar manner, the minimum number of sentences read by a single male speaker is 61 and the maximum is 151. Out of the total speakers, 120 are recorded in an office environment and the remaining 13 are recorded in a studio. All speakers are recorded at 16 MHz sampling rate using Audacity 1.3 Beta (Unicode). Before parameterization, we amplified the weak speeches and attenuated the constant background noises.



	a	u	i	A	E	e	o
h	ሀ [3211]	ሁ [3451]	ሂ [398]	ሃ [3211]	ሄ[1032]	ህ [4716]	ሆ[3362]
l	ለ [16185]	ሉ [3295]	ሊ [2622]	ላ [8469]	ሌ[1519]	ል [17549]	ሎ[2032]
m	መ [15533]	ሙ[1488]	ሚ [7797]	ማ[9067]	ሜ[935]	ም [15233]	ሞ[2256]
s	ሠ [7533]	ሡ [1300]	ሢ [2055]	ሣ [4097]	ሤ[772]	ሥ [21007]	ሦ[1457]
r	ረ [7684]	ሩ [3049]	ሪ [4093]	ራ [8267]	ሬ[2159]	ር [17055]	ሮ[3252]
q	ቀ [4751]	ቁ [1412]	ቂ [618]	ቃ [2443]	ቄ[293]	ቅ [3725]	ቆ[1076]
b	በ [19453]	ቡ [1941]	ቢ [1864]	ባ [9040]	ቤ[1327]	ብ [8023]	ቦ[1004]
t	ተ [17075]	ቱ [2859]	ቲ [1213]	ታ [6755]	ቴ[1078]	ት [30462]	ቶ[3087]
c	ቸ [6611]	ቹ [1137]	ቺ [135]	ቻ [1543]	ቼ[148]	ች [11908]	ቾ[263]
n	ነ [11709]	ኑ [1955]	ኒ [1347]	ና [12539]	ኔ[1169]	ን [40727]	ኖ[1555]
N	ኘ [662]	ኙ [846]	ኚ [80]	ኛ [2181]	ኜ[72]	ኝ [1120]	ኞ[799]
H	አ [19153]	ኡ [173]	ኢ [4880]	ኣ [19153]	ኤ[1911]	እ [13879]	ኦ[725]
k	ከ [7660]	ኩ [1286]	ኪ [455]	ካ [3742]	ኬ[408]	ክ [5228]	ኮ[2006]

Fig.2: Frequency of some of the syllables in the training set

4. Labeling

The UASR toolkit, developed by TU Dresden will be used to build the Amharic speech recognizer. This toolkit requires that some part of the speech corpus must be manual labeled in order to initialize the system. Therefore, a speech set of 143.96 minutes duration is manually labeled using wavesurfer version 8.4.2.4.

A script is written using a programming language called dlabpro, which is developed by TU Dresden. This script reads the transcription file of the speeches line by line and generates a separate dummy file corresponding to each line. The output file generated by this script has three columns: the first column is the starting time of a syllable, the second is the final time of a syllable, and the last column is name of a syllable. The beginning and final times in these files are not correct. They are only there as a placeholder. They are replaced by the actual values by executing the HMM.xtp script of UASR. HMM.xtp sequentially reads the dummy files and replaces the start and final syllable times with actual values.

5. The lexical and language models

A syllabic lexical model consisting of more than 227,773 Amharic words has been prepared. It is an alternative lexical model in that more than one pronunciation is provided for some of the words. The word ‘sl’ is included in the dictionary and mapped to long silences at the beginning and end of the speeches. The short-pause is modeled by adding “sp” at the end of every word pronunciation.

In order to build our finite-state-automata based tri-gram language model, a text corpus of more than 10,000 pages has been collected from different Amharic news websites. The texts are originally written using the Ge’ez script. But we have replaced each Ge’ez character by two Roman characters as UASR can only process texts written using Roman letters.

We used the free Toolkits called ‘GRM Library’ and ‘FSM Toolkit’ from AT&T to build the language model. First the text corpus is converted into a format called ‘FAR archive’ using the farcompilestrings command of the FSM Toolkit. This command reads the text corpus and replaces each line in the corpus by a finite-state-automaton in which each word in the sentence is the label of a transition from one state to another. Once the FAR archive is generated, it is passed to a pipelined instruction to generate a Katz backoff trigram language mode as shown below:

```
farcompilestrings -i word_list.wl –u "<u>" corpus.txt | grmcount –n 3 –s <s>
-f </s> –i word_list.wl | grmmake | grmshrink > language_model.fsm
```

6. The Acoustic model

We use syllables instead of phones as the fundamental sub-word units to build the acoustic model. We opted for this method for two important reasons. Firstly, during the manual labeling procedure, it is far easier to accurately identify boundaries of syllables than phonemes; secondly, the number of syllables in the Amharic language is very small. Unlike other languages, there are only 233 syllables in Amharic.

At the moment we used only the manually labeled speeches to train the system. The current syllable recognition accuracy is only 41%. We expect to significantly improve this accuracy when we use all the speeches we have after automatically labeling them.

7. References

[1] Solomon Berhanu, “Isolated Amharic consonant-vowel syllable recognition: An experiment using the Hidden Markov Model,” Msc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia, 2001.

[2] Kinfe Tadesse, “Sub-word based Amharic speech recognizer: An experiment using Hidden Markov Model (HMM),” MSc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia, June 2002.

[3] Martha Yifiru, “Automatic Amharic speech recognition system to command and control computers,” MSc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia, 2003.

[4] Zegaye Seifu, “HMM based large vocabulary, speaker independent, continuous Amharic speech recognizer,” MSc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia, 2003.

[5] Solomon Tefera, “Automatic Speech Recognition for Amharic”, PhD Dissertation, Hamburg University, 2005.