An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition

Solomon Teferra Abate, Wolfgang Menzel, Bairu Tafila

University of Hamburg, Faculty of Informatics

{solomon,menzel}@nats.informatik.uni-hamburg.de

1. The Amharic Language

Amharic:

- is fficial language of Ethiopia.
- is a Semitic language family.
- has the largest number of speakers after Arabic - 22.5 million.
- has five dialectical variations: Addis Ababa, Gojjam, Gonder, Wollo, and Menz.
- has special features. For example: glottalized plosives ጵ, ጥ, ጽ, ፍ, and ቅ.
- has rich morphology -> many word forms.

Phonetics

Amharic has a set of 38 phones, seven vowels and thirty-one consonants.

Consonants

Manner of Art/n	Voicing	Place of Articulation				
		Lab	Dent	Pal	Vel	Clo
Stops	Voiceless	ፕ[p]	ት[t]	ች[tʃ]	ክ[k]	ኣ[ʔ]
	Voiced	ብ[b]	ድ[d]	ጅ[dʒ]	ግ[g]	
	Glottalized	ጵ[pʼ]	ጥ[tʼ]	ጭ[tʃʼ]	ቅ[q]	
	Rounded				ኵ [k^w], ጕ[g^w], ቍ[q^w]	
Fricatives	Voiceless	ፍ[f]	ስ[s]	ሽ[ʃ]		ሀ[h]
	Voiced		ዝ[z]	ዥ[ʒ]		
	Glottalized		ጽ[sʼ]			
	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]	

Vowels

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

The writing system

- Amharic writing system is phonetic.
- Amharic orthography is syllabic, because
  - there is, more or less, a one-to-one correspondence between the sounds of CV syllable and the graphemes.
- Amharic orthography consist of 276 symbols, including the redundant ones.
- Amharic has 234 distinct CV syllable sounds.

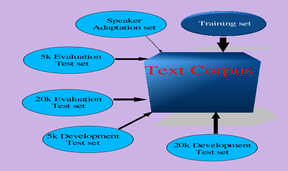
Used graphemes

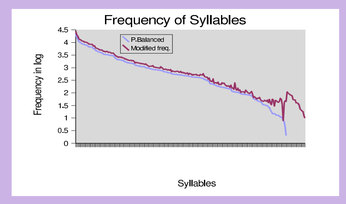
		ə	u	i	A	E	e	o	ua	ui	uA	uE	ue
h	h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ	ኈ	ኍ	ኋ	ኌ	ኈ
l	l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ			ሏ		
m	m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ			ሟ		
r	r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ			ሯ		
s	s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ			ሷ		
S	^s	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ			ሿ		
q	q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቈ	ቍ	ቋ	ቌ	ቈ
b	b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ			ቧ		
v	v	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ					
t	t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ			ቷ		
c	^c	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ			ቿ		
n	n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ			ኗ		
N	^n	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ			ኟ		
H	ʼ	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ					
k	k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኰ	ኵ	ኳ	ኴ	ኰ
K	^k	ኸ											
w	w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ					
z	z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ			ዟ		
Z	^z	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ			ዧ		
y	y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ					
d	d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ			ዷ		
D	^d	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ					
g	g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ	ጐ	ጒ	ጓ	ጔ	ጐ
T	t	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ			ጧ		
C	^C	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ	ጨ		ጯ		
P	p	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ					
x	c	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ			ፇ		
f	f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ			ፏ		
p	p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ					

2. Preparation of Amharic Speech Corpus

Text Corpus Preparation

- It is a read speech corpus.
- Text was aquired from EthioZena website.
- Preprocessing:
  - 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.





Recording of the speech

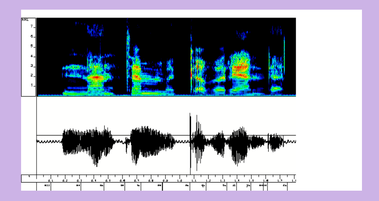
We have done the recording using a laptop and a noise canceling micraphone.

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				
18-23	10	3	4	
24-28	6	1		
Total	16	4	4	
Grand Total	56	44	14	10

- Size of the corpus:
  - 20 hours of training speech
  - 100 speakers
  - 10850 sentences
  - Vocabulary of training speech is 28666
- The training speech covers 233 of the Amharic CV syllables

Segmentation of the speech

Segmentation is done semi-automatically at word- and syllable-level.



3. Availability

- Used for our research of developing automatic speech recognizer for Amharic.
- Part of it is used by students of Addis Ababa University.
- It will be made available through a third party.

4. Acknowledgments

Daniel Yacob who made the archive of EthioZena available to us. He also provided his SERA to Ethiop (namely g2) conversion tool.
DAAD (Deutscher Akademischer Austauschdienst) financed the project.

God bless you