Experimental detection of vowel pronunciation variants in Amharic

Thomas Pellegrini and Lori Lamel

LIMSI-CNRS, BP133
91403 Orsay cedex, FRANCE
{thomas.pellegrini, lamel}@limsi.fr

Abstract

The pronunciation lexicon is a fundamental element in an automatic speech transcription system. It associates each lexical entry (usually a grapheme), with one or more phonemic or phone-like forms, the pronunciation variants. Thorough knowledge of the target language is a priori necessary to establish the pronunciation baseforms and variants. The reliance on human expertise can pose difficulties in developing a system for a language where such knowledge may not be readily available. In this article a speech recognizer is used to help select pronunciation variants in Amharic, the official language of Ethiopia, focusing on alternate choices for vowels. This study is carried out using an audio corpus composed of 37 hours of speech from radio broadcasts which were orthographically transcribed by native speakers. Since the corpus is relatively small for estimating pronunciation variants, a first set of studies were carried out at a syllabic level. Word lexica were then constructed based on the observed syllable occurences. Automatic alignments were compared for lexica containing different vowel variants, with both context-independent and context-dependent acoustic models sets. The *variant2+* measure proposed in (Adda-Decker and Lamel, 1999) is used to assess the potential need for pronunciation variants.

1. Introduction

For large vocabulary speech recognition systems, the development of pronunciation lexica typically requires human knowledge and manual intervention. For languages with a close to phonemic writing system, pronunciation dictionary can be created using grapheme-to-phoneme rules. While these can provide initial base forms, one means of improving system performance is to add pronunciation variants to the lexicon. However, to exhaustively determine pronunciation rules can be very difficult even for someone who has good knowledge in the target language. All languages have exceptions and a lot of time is needed to enumerate them. In Brazilian Portuguese for example, there is sometimes a reduction of the diphthong /ej/ to /e/ (brasileiro is pronounced [bɹazi'leɹu] and rarely [bɹazilejɹu]). There are other words that don't suffer this reduction: peito and lei, for example, which are pronounced [pejtu] and [lej]. The realization of the vowel 'o'in American English and British English can be more or less diphthongized depending upon the local accent.

With the availability of large corpora of transcribed speech alternative approaches have been proposed to introduce pronunciation variants. Various studies have described attempts to automatically determine pronunciation variants, such as (Cohen, 1989; Riley and Ljojle, 1996). One approach is to use rules to over generate pronunciations in a preliminary working dictionary and validate their selection on a lot of data (Adda-Decker and Lamel, 1999). In (Adda-Decker et al. , 2005), an automatic speech recognition system is used as a linguistic tool to investigate syllabic structures and their variation in spontaneous French. This paper reports on a corpus-based method to help select pronunciation variants for the Amharic language.

There are various recent studies on speech recognition and speech processing for Amharic (Abate et al. , 2005; Seid and Gamback, 2005; Eyassu and B. Gamback, 2005), a new resource web portal for Amharic corpora has also been created. [1] Compared to other languages for which models and systems have been developed at LIMSI (Lamel and Gauvain, 2002), the Amharic audio corpus is quite small, containing a total of 240k words with 50k distinct lexemes. Since Amharic is a syllabic language, a syllable level representation was used to first identify syllabic variants that could then be used to generate pronunciation variants at the word level. Context-independent and context-dependent acoustic models have been used with two representations:

- a syllabotactic representation, in order to detect the potential pronunciation variants;
- a whole word representation to validate the variants previously hypothesized.

2. Brief presentation of Amharic

Amharic is the official language of Ethiopia and has about 14 million speakers (source: omniglot.com). Although it is a Semitic language like Hebrew and Arabic, its writing, developed from the Ethiopian classical language Ge'ez, is a syllabic left-to-right script. Amharic has 34 basic symbols, for which there are 7 vocalizations referred to as the seven orders: /ɛ/, /u/, /i/, /a/, /e/, /ə/ and /o/. The basic symbols are modified in a number of different ways to indicate the various vowels. 85% of these represent a CV sequence (C for consonant and V for vowel), one symbol representing the complex sound /ts/ and the remaining a CwV sequence (w is a semi-consonant).
Amharic has different levels of spelling conventions (Yacob, 2003). At the basic level, words are written as they are spoken, with little care about consistent spelling. Some of the spelling variations arise from the presence of homophone symbols. For example, there are 3 basic symbols corresponding to the sound /h/. A word can have numerous different spellings but according to Yacob (2003), the level of conformance in newspapers and literature is relatively high.

[1] http://corpora.amharic.org/

3. Motivation

Since there is a one-to-one correspondence between the transliteration character set and the phone set, a first pronunciation lexicon was generated by simply repeating the full transliterated words as both the lexical entry and the phonemic form. The 240 Amharic symbols were mapped to a set of 33 characters corresponding to phonemes. This mapping or transliteration not only reduces the number of graphemes to be used but also eliminates the problem of homophonic forms. This differs from a grapheme-based approach since the homophones are merged in one entry.
The following is an example Amharic (a few introductory words of the radio Medhin news) sentence taken from the audio corpus, followed by its transliteration:

የኢትዮጵያ መድህን ድምፅ ራዲዮ
jE?itxjoPxja mEdxhxnx dxmxtsx radijo

Analyzing the first alignments carried out with the full form pronunciation dictionary, it was observed that very often schwas were not pronounced. This caused large misalignments for some of the segments. A second word based lexicon was constructed in which schwas were allowed to be optional, which on average improved the segmentation quality. Looking at alignments with this dictionary a fair number of (unforeseen) vowel alternates were observed. In order to further explore alternative vowel pronunciations, we decided to use syllable alignments since these could provide a better generalization capability than word alignments. Alternate syllable pronunciations were explored, and the pronunciations selected by the system during alignment were extracted and that most frequent vowel alternates were used to build new full-word lexica.

4. The syllabotactic representation

For these studies the lexicon is comprised of a simple list of all the syllables in Amharic. In addition to the default baseline pronunciation, the vowel was allowed to be optional and could be replaced by any other vowel.

4.1. The syllable lexicon

The phone set is comprised of 33 phones including the vowels, with 3 additional phones (for silence, breath and hesitation). The Cw sequences are modeled by separate units for C and w. For a first syllabic representation, the lexicon is simply a list of all the syllables. The information of the syllable position is preserved by adding a sign (here a underscore) on the right and/or left of each syllable to ensure the ability to unambiguously recombine the syllables into words. Table 1 gives an example of the syllable representation for the word "bEdemokxrasi" ("ቦደሞክረሲ"):

Word	*Syllabotactic form*
ቦደሞክረሲ	ቦ_ _ደ_ _ሞ_ _ክ_ _ረ_ _ሲ
bEdemokxrasi	bE_ _de_ _mo_ _kx_ _ra_ _si

Table 1: Example of a word and its transliteration

Two entries of syllables starting with 'd' are shown in Table 2 along with the automatically generated alternate pronunciations for vowels The baseline pronunciations for these syllables are shown in bold.

_dE	dx	du	do	di	de	da	**dE**	d
_dE_	dx	du	do	di	de	da	**dE**	d
dE_	dx	du	do	di	de	da	**dE**	d
_da	dx	du	do	di	de	**da**	dE	d
_da_	dx	du	do	di	de	**da**	dE	d
da_	dx	du	do	di	de	**da**	dE	d

Table 2: Sample entries in the syllabic lexicon. The baseline pronunciations are shown in bold.

4.2. Experimental results

Table 3 shows a confusion matrix (in %) when using the syllabic lexicon for alignment. The columns the graphemic forms given in the reference transcription and the rows are the phonemic forms selected during forced alignment. The first row (#Occ) gives the number of occurrences of the vowel in the corpus, and the last row gives the percentage of times the vowel is deleted. It can been seen that schwa deletion (represented by x) is very frequent, over 60% being deleted. There are no other dominant schwa substitutions with other vowel. The /E/ seems to be rather unstable as less than 50% are aligned as /E/. The most common substitution is from /E/ to /a/ (10.6%) and the deletion rate is over 20%. Three pairs of vowels with high substitution rates are shown in bold in the table: /E/–/a/; /e/–/i/; /o/–/u/. It is thought that most early Semitic languages had only three vowels (plus the schwa). The Arabic language nominally has only three vowels /a, u, i/ which correspond to the predominant confusion classes in Amharic.

	Graphemes						
	E	a	e	i	o	u	x
#Occ	290k	156k	22k	51k	41k	40k	363k
ɛ	**47.4**	**7.1**	6.1	1.5	5.7	1.9	4.8
a	**10.6**	**78.0**	1.4	1.0	2.0	0.6	1.5
e	6.7	0.7	**64.0**	**11.0**	1.5	0.9	2.7
i	1.9	0.3	**11.2**	**53.8**	1.4	1.8	4.2
o	4.4	1.0	0.9	0.8	**67.1**	**14.8**	1.8
u	1.4	0.2	0.6	1.2	**10.7**	**57.1**	3.5
ə	5.7	0.5	2.3	5.1	3.7	5.7	**20.0**
Del	21.8	12.2	13.5	25.5	8.1	17.2	**61.4**

Table 3: Vowel confusion matrix using syllable lexicon.

While Table 3 gives a global view of the vowel characteristics as seen by the system, allowing all frequent substitutions as pronunciation variants would introduce too many forms at the word level. In order to have a finer view, vowel confusions were determined for each individual syllable onset (C or Cw). Given the syllabotactic representation which allows syllables to be recombined into words it would also be possible to consider the right context, but this has not been done yet.
Table 4 gives the vowel substitution and deletion percentages for syllables with the onset 'b'. The columns correspond to the graphemic form, with the number of occurrences given in the first row and the different phonemic forms in the following rows. In the first column it can be seen that almost 55% of syllable 'bE'are selected as /bɛ/, 17% of /ɛ/ are deleted and 10.6% are substituted with /bo/. These results differ somewhat from the global tendencies

	Graphemes						
	bE	ba	be	bi	bo	bu	bx
#Occ	30.5k	12.3k	2.2k	2.7k	1.3k	2.5k	12.8k
ɛ	**54.8**	3.0	4.4	1.1	4.0	1.1	7.0
a	9.1	**87.4**	1.0	1.4	1.2	0.5	1.6
e	2.1	0.2	**76.4**	**11.1**	0.6	0.3	2.5
i	0.6	0.3	**11.6**	**53.8**	0.4	0.8	4.8
o	**10.6**	2.3	1.5	1.1	**79.7**	**21.1**	5.9
u	2.2	0.1	0.6	1.7	10.3	**61.9**	8.9
ə	3.7	0.1	1.7	2.7	0.6	1.6	**13.8**
Del	**16.9**	**6.6**	2.8	**27.1**	3.2	12.7	**55.5**

Table 4: Confusion matrix for 'b'syllables.

shown in the Table 3 where the largest substitution for /ɛ/ was /a/. For the syllable 'be', the vowel is relatively stable (76.3% /e/, 2.8% deletions) and the largest confusion is with /i/, as observed globally for this vowel. It can be noted that two other vowels (a and o) are also quite stable with low deletion rates and substitution rates of about 20%.

5. Word level

Using the insight gained at the syllable level, experiments were carried out to assess the need for pronunciation variants at the word level.

5.1. The word lexicon

Rules for vowel substitution and deletion were determined for each syllable using confusion matrices similar to the example given in Table 4. Different pronunciation lexicons have been generated but only one, which has only the most frequent variant for each syllable, is used in these experiments. For example, for the 'bE'syllable, the alternate pronunciation is with a deletion of the vowel. For the 'be'syllable, the alternate substitutes the vowel /i/ for the vowel /e/. With the baseline form and one variant for each syllable, each lexeme has 2^N potential alternates where N is the number of syllables in the word. Figure 1 shows the number of distinct words in the audio corpus as a function of word length in phonemes. The most common word length is 10 phonemes (corresponding to five CV syllables), mainly arising from the gluing of affixes for articles, demonstratives, plural marks, etc... For 5-syllable words, there are $2^5 = 32$ variants in the lexicon. The choice of including one variant per syllable ensures that short words (the ten most frequent words in the corpus are bisyllabic) have variants while avoiding an explosion of the number of variants for long words. As in other languages, frequent words are a priori expected to be subject to pronunciation reductions, so it seems important to provide variants for them.

Figure 2 shows the evolution of the average word length as a function of the word frequency rank in the audio transcriptions. The less frequent words have an average length that is almost the double that of the most frequent words. The average number of potential pronunciation variants increases with the word frequency rank.

Some example lexical entries are shown in Table 5. In the phonemic forms, vowels into braces are optional and vowel pairs in brackets are interchangeable.

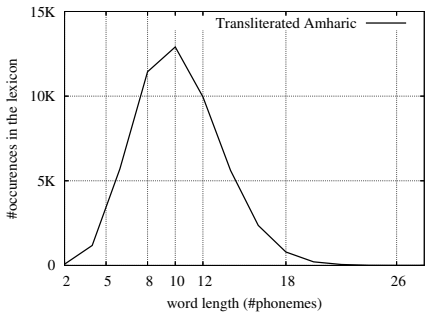

Figure 1: Lexeme distribution as a function of word length in phonemes (50.3k distinct words).

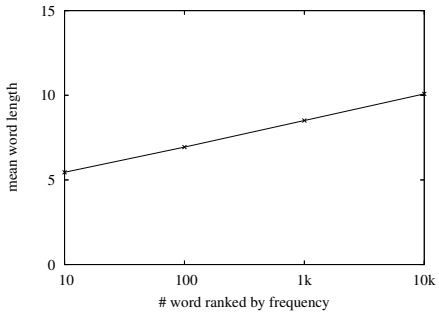

Figure 2: Average word length (# phonemes) vs word frequency rank in the transcriptions.

Lexical Entry	*Phonemic form*
nEwx	n{E}w{x}
mEto	m{E}t[ou]
jEdemokxrasi	j{E}d[ei]m[ou]k{x}r{a}s{i}

Table 5: Sample entries in the word lexicon.

5.2. Experimental results

To measure the need for pronunciation variants, the *variant2+* rate was proposed in (Adda-Decker and Lamel, 1999). The *variant2+* rate is the percentage of words aligned with variants that are not the "main"phonemic form. In our case, the main form is the phonemic form identical to the graphemic form, i.e. all phonemes are required.

Alignments were made with different acoustic models sets: 36 context-independent models (CI models), and models with 100, 300, 1k and 7.2k triphone contexts (CD models). The same audio corpus is used to compute the *variant2+* rate as was used to train the acoustic models, which might lead to an underestimate since the acoustic models may implicitly model some of the variation. Table 6 shows results with 100 CD models for two frequent bisyllablic words. For each word the frequency rank, number of occurrences (#occ), number of aligned occurrences (#align), variant2+ rate, and different phonemic forms (phones) with their percentages of the aligned occurrences are given. The most frequent word in the corpus, nEwx, has all of its phonemes pronounced in only about 10% of the time. The most frequent realization chosen by the aligner is without the final schwa (81.5% of the occurrences). Not surprisingly the last form with both vowels deleted is never selected by the

system. The second word has a different comportment in that about half of the word occurrences are aligned with the baseform. The pronunciation rules used for this word allow the deletion of /ɛ/ and the substitution of /o/ by /u/. The second most frequent phonemic form is /mɛtu/ (36.1%). The forms with the deleted first vowel have low selection rates (6.4% and 7.0%).

Word	*Rank*	*#Occ*	*#Align*	*Variant2+*	*Phones*	%
nEwx	1	3044	2964	90.3%	nɛwə	9.7
					nwə	8.7
					nɛw	81.5
					nw	0.0
mEto	7	803	783	49.5%	mɛto	50.4
					mto	6.4
					mɛtu	36.1
					mtu	7.0

Table 6: *Variant2+* rates and percentages of pronunciation variants for two frequent words.

These two examples reflect some global tendencies of the alignments: words with schwas are mainly aligned with reduced forms (schwa deletion). Words without schwas are aligned with fewer pronunciation variants. The average variant2+ rate for words with schwas is 86.9% compared to 51.3% for words without schwas.

Figure 3 shows the variant2+ rate as a function of the word frequency rank for different acoustic models sets. Four points are given for each curve: for word frequency ranks of 10, 100, 1k and 10k. These correspond to the average of the variant2+ rates for words with a rank within an interval centered at these values. The interval length varies with its center. Table 7 gives the size of the interval, the average number of occurrences (AvOcc), the average length of the words (wl), and the percentage of words with schwas (%schwa) in the interval for each value.

Rank	*Rank interval*	*AvOcc*	*wl*	*%schwa*
10	4-16	710	5.4	72%
100	74-126	220	6.9	80%
1k	899-1101	30	8.5	82%
10k	9499-10501	3	10.0	85%

Table 7: Intervals used to measure the *variant2+* rate.

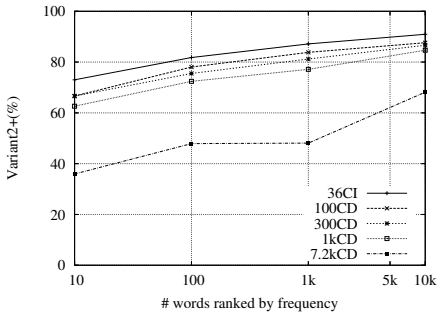

Figure 3: *Variant2+* vs word frequency rank.

Since context-dependent models implicitly capture acoustic variations, the larger the number of modeled contexts, the less of a need for pronunciation variants during the alignment stage. This can be seen in from the curves where the 36 CI models have the highest variant2+ rate and the 7.2k CD models has the lowest one. The variant2+ rates are seem to increase with word frequency rank for all model sets. This can be explained with various factors: the increase in the average word length with the frequency rank and the increase of the number of words containing schwas. The average word length doubles from the most frequent words (a rank around 10) to the least frequent words (a rank around 10k). 85% of the words in the interval around 10k have at least one schwa and these words are mostly aligned with variants.

6. Summary

In this paper a syllabotactic representation has been used as a means to determine potential pronunciation variants. A syllable-based lexicon allowing all potential vowel substitutions and deletions was confronted with a transcribed audio corpus. The most frequent variants for each syllable onset were used in build a word-based lexicon. The inclusion of pronunciation variants during forced alignment seems to improve the quality of the alignments and the likelihood of the acoustic models. The word alignments were carried out to provide counts for the variants and most frequent variants were selected for the speech recognizer lexicon. Speech recognition experiments using different pronunciation lexica are underway. This approach has been used for a syllabic language but could be applied to other languages as well, and is of particular interest when only a small audio corpus and limited amount of linguistic knowledge are readily available.

7. References

S.T. Abate, W. Menzel, and B. Tafila. 2005. An amharic speech corpus for large vocabulary continuous speech recognition. In *Proc. Interspeech*, Lisbon.

M. Adda-Decker and L. Lamel. 1999. Pronunciation variants across system configuration, language and speaking style. *Speech Communication*, 29:83–98.

M. Adda-Decker, P. Boula de Mareuil, G. Adda, and L. Lamel. 2005. Investigating syllabic structures and their variation in spontaneous french. *Speech Communication*, 46:119–139.

M. Cohen. 1989. *Phonological structures for speech recognition*. Ph.D. thesis, U.Ca. Berkeley.

S. Eyassu and B.Gamback. 2005. Classifying amharic news texts uing self-organizing maps. In *ACL05 Workshop on computational Approaches to Semitic Languages*, Ann Arbor, Michigan.

L. Lamel and J.L. Gauvain. 2002. Automatic processing of broadcast audio in multiple languages. In *Eusipco02*, Toulouse.

M.D. Riley and A. Ljojle, 1996. *Automatic Generation of Detailed Pronunciation Lexicons*, pages 285–301.

H. Seid and B. Gamback. 2005. A speaker independent continuous speech recognizer for amharic. In *Proc. Interspeech*, Lisbon.

D. Yacob. 2003. Application of the double metaphone algorithm to amharic orthography. In *International Conference of Ethiopian Studies XV*.