Development of an Amharic Text-to-Speech System Using Cepstral Method

Tadesse Anberbir
ICT Development Office, Addis Ababa University, Ethiopia
tadanberbir@gmail.com

Tomio Takara
Faculty of Engineering, University of the Ryukyus, Okinawa, Japan
takara@ie.u-ryukyu.ac.jp

Abstract

This paper presents a speech synthesis system for Amharic language and describes and how the important prosodic features of the language were modeled in the system. The developed Amharic Text-to-Speech system (AmhTTS) is parametric and rule-based that employs a cepstral method. The system uses a source filter model for speech production and a Log Magnitude Approximation (LMA) filter as the vocal tract filter. The intelligibility and naturalness of the system was evaluated by word and sentence listening tests respectively and we achieved 98% correct-rates for words and an average Mean Opinion Score (MOS) of 3.2 (which is categorized as good) for sentences listening tests. The synthesized speech has high intelligibility and moderate naturalness. Comparing with previous similar study, our system produced considerably similar quality speech with a fairly good prosody. In particular our system is mainly suitable for building new languages with little modification.

1 Introduction

Text-to-Speech (TTS) synthesis is a process which artificially produces synthetic speech for various applications such as services over telephone, e-document reading, and speaking system for handicapped people etc.

The two primary technologies for generating synthetic speech are concatenative synthesis and formant (Rule-based) synthesis methods. Concatenative synthesis produces the most natural-sounding synthesized speech. However, it requires a large amount of linguistic resources and generating a various speaking style is a challenging task. In general the amount of work required to build a concatenative system is enormous. Particularly, for languages with limited linguistic resources, it is more difficult. On the other hand, formant synthesis method requires small linguistic resources and able to generate various speaking styles. It is also suitable for mobile applications and easier for customization. However, this method produced less natural-sounding synthesized speech and the complex rules required to model the prosody is a big problem.

In general, each method has its own strengths and weaknesses and there is always a tradeoff. Therefore, which approach to use will be determined by the intended applications, the availability of linguistic resources of a given language etc. In our research we used formant (rule-based) synthesis method because we are intending to prepare a general framework for Ethiopian Semitic languages and apply it for mobile devices and web embedded applications.

Currently, many speech synthesis systems are available mainly for 'major' languages such as English, Japanese etc. and successful results are obtained in various application areas. However, thousands of the world's 'minor' languages lack such technologies, and researches in the area are quite very few. Although recently many localization projects (like the customization of Festvox[1]) are being undergoing for many languages, it is quite inadequate and the localization process is not an easy task mainly because of the lack of linguistic resources and absence of similar works in the area. Therefore, there is a strong demand for the development of a speech synthesizer for many of the African minor languages such as Amharic.

Amharic, the official language of Ethiopia, is a Semitic language that has the greatest number of speakers after Arabic. According to the 1998 census, Amharic has 17.4 million speaker as a mother thong language and 5.1 million speakers as a second language. However, it is one of the

[1] Festvox is a voice building framework which offers general tools for building unit selection voices for new languages.

Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009, pages 46–52,
Athens, Greece, 31 March 2009.

least supported and least researched languages in the world. Although, recently, the development of different natural language processing (NLP) tools for analyzing Amharic text has begun, it is often very far comparing with other languages (Alemu et al., 2003). Particularly, researches conducted on language technologies like speech synthesis and the application of such technologies are very limited or unavailable. To the knowledge of the authors, so far there is only one published work (Sebsibe, 2004) in the area of speech synthesis for Amharic. In this study they tried to describe the issues to be considered in developing a concatenative speech synthesizer using Festvox and recommended using syllables as a basic unit for high quality speech synthesis.

In our research we developed a syllabic based TTS system with prosodic control method which is the first rule-based system published for Amharic. The designed Amharic TTS (AmhTTS) is parametric and rule-based system that employs a Cepstral method and uses a Log Magnitude Approximation (LMA) filter. Unlike the previous study, Sebsibe (2004), our study provides a total solution on prosodic information generation mainly by modeling the durations. The system is expected to have a wide range of applications, for example, in software aids to visually impaired people, in mobile phones and can also be easily customized for other Ethiopian languages.

2 Amharic Language's Overview

Amharic (**አማርኛ**) is a Semitic language and it is one of the most widely spoken languages in Ethiopia. It has its own non Latin based syllabic script called "Fidel" or "Abugida". The orthographic representation of the language is organized into orders (derivatives) as shown in Fig.1. Six of them are CV (C is a consonant, V is a vowel) combinations while the sixth order is the consonant itself. In total there are 32 consonants and 7 vowels with 7x32= 224 Syllables. But since there are redundant sounds that represent the same sounds, the phonemes are only 28 (see the Appendix).

1st	2nd	3rd	4th	5th	6th	7th
C/e/ ቸ	C/u/ ቹ	C/i/ ቺ	C/a/ ቻ	C/ie/ ቼ	C ች	C/o/ ቾ

Figure 1: Amharic Syllables structure

Like other languages, Amharic also has its own typical phonological and morphological features that characterize it. The following are some of the striking features of Amharic phonology that gives the language its characteristic sound when one hears it spoken: the weak, indeterminate stress; the presence of glottalic, palatal, and labialized consonants; the frequent gemination of consonants and central vowels; and the use of an automatic helping vowel (Bender et al., 1976).

Gemination in Amharic is one of the most distinctive characteristics of the cadence of the speech, and also caries a very heavy semantic and syntactic functional weight (Bender and Fulass, 1978). Amharic gemination is either lexical or morphological. Gemination as a lexical feature cannot be predicted. For instance, **አለ** may be read as *alä* meaning 'he said', or *allä* meaning 'there is'. Although this is not a problem for Amharic speakers, it is a challenging problem in speech synthesis. As a morphological feature gemination is more predictable in the verb than in the noun, Bender and Fulass (1978). However, this is also a challenging problem in speech synthesis because to automatically identify the location of geminated syllables, it requires analysis and modeling of the complex morphology of the language. The lack of the orthography of Amharic to show geminates is the main problem. In this study, we used our own manual gemination mark (') insertion techniques (see Section 3.3.1).

The sixth order syllables are the other important features of the language. Like geminates, the sixth order syllables are also very frequent and play a key role for proper pronunciation of speech. In our previous study, (Tadesse and Takara, 2006) we found that geminates and sixth order syllables are the two most important features that play a key role for proper pronunciation of words. Therefore, in our study we mainly consider these language specific features to develop a high quality speech synthesizer for Amharic language.

3 AmhTTS System

Amharic TTS synthesis system is a parametric and rule based system designed based on the general speech synthesis system. Fig.2. shows the scheme of Amharic speech synthesis system. The design is based on the general speech synthesis system (Takara and Kochi, 2000). The system has three main components, a text analysis subsystem, prosodic generation module and a speech synthesis subsystem. The following three sub-sections discuss the details of each component.

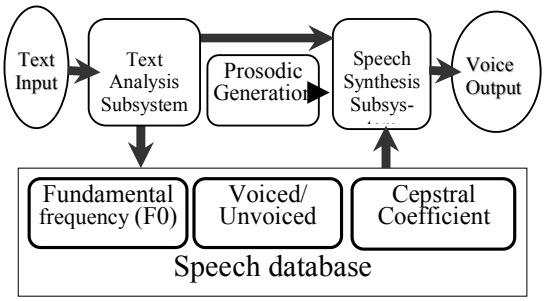

Figure 2: Amharic Speech Synthesis System

3.1 Text Analysis

The text analysis subsystem extracts the linguistic and prosodic information from the input text. The program iterates through the input text and extracts the gemination and other marks, and the sequences of syllables using the syllabification rule. The letter-to-sound conversion has simple one-to-one mapping between orthography and phonetic transcription (see Apendix). As defined by (Baye, 2008; Dawkins, 1969) and others, Amharic can be considered as a phonetic language with relatively simple relationship between orthography and phonology.

3.2 Speech Analysis and Synthesis systems

First, as a speech database, all Amharic syllables (196) were collected and their sounds were prepared by recording on digital audio tape (DAT) at a 48 kHz sampling rate and 16-bit value. After that, they were down-sampled to 10 kHz for analyzing. All speech units were recorded with normal speaking rate.

Then, the speech sounds were analyzed by the analysis system. The analysis system adopts short-time cepstral analysis with frame length 25.6 ms and frame shifting time of 10 ms. A time-domain Hamming window with a length of 25.6 ms is used in analysis. The cepstrum is defined as the inverse Fourier transform of the short-time logarithm amplitude spectrum (Furui, 2001). Cepstral analysis has the advantage that it could separate the spectral envelope part and the excitation part. The resulting parameters of speech unit include the number of frames and, for each frame, voiced/unvoiced (V/UV) decision, pitch period and cepstral coefficients c[m], $0 \leq m \leq 29$. The speech database contains these parameters as shown in fig.2.

Finally, the speech synthesis subsystem generates speech from pre-stored parameters under the control of the prosodic rules. For speech synthesis, the general source-filter model is used as a speech production model as shown in fig.3. The synthetic sound is produced using Log Magnitude Approximation (LMA) filter (Imai, 1980) as the system filter, for which cepstral coefficients are used to characterize the speech sound.

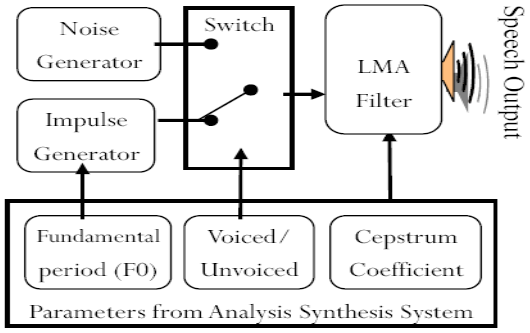

Figure 3: Diagram of Speech Synthesis Model

The LMA filter presents the vocal tract characteristics that are estimated in 30 lower-order quefrency elements. The LMA filter is a pole-zero filter that is able to efficiently represent the vocal tract features for all speech sounds. The LMA filter is controlled by cepstrum parameters as vocal tract parameters, and it is driven by fundamental period impulse series for voiced sounds and by white noise for unvoiced sounds. The fundamental frequency (F0) of the speech is controlled by the impulse series of the fundamental period. The gain of the filter or the power of synthesized speech is set by the 0th order cepstral coefficient, c [0].

3.3 Prosody Modeling

For any language, appropriate modeling of prosody is the most important issue for developing a high quality speech synthesizer.

In Amharic language segments duration is the most important and useful component in prosody control. It is shown that, unlike English language in which the rhythm of the speech is mainly characterized by stress (loudness), rhythm in Amharic is mainly marked by longer and shorter syllables depending on gemination of consonants, and by certain features of phrasing (Bender et al., 1976). Therefore it is very important to model the syllables duration in AmhTTS system. In this paper we propose a new segmental duration control methods for synthesizing a high quality speech. Our rule-based TTS system uses a compact rule-based prosodic generation method in three phases:

- modeling geminated consonants duration

- controlling of sixth order syllables duration
- assignment of a global intonation contour

Prosody is modeled by variations of pitch and relative duration of speech elements. Our study deals only with the basic aspects of prosody such as syllables duration and phrase intonation. Gemination is modeled as lengthened duration in such a way geminated syllables are modeled on word level. Phrase level duration is modeled as well to improve prosodic quality. Prosodic phrases are determined in simplified way by using text punctuation. To synthesize F0 contour Fujisaki pitch model that superimpose both word level and phrase level prosody modulations is used (Takara and Jun, 1988).

The following sub-sections discuss the prosodic control methods employed in our system.

3.3.1 Gemination rule

Accurate estimation of segmental duration for different groups of geminate consonants (stops, nasals, liquids, glides, fricatives) will be crucial for natural sounding of AmhTTS system. In our previous study, Tadesse and Takara (2006), we studied the durational difference between singletons vs. geminates of contrastive words and determined the threshold duration for different groups of consonants. Accordingly the following rule was implemented based on the threshold durations we obtained in our previous study.

The gemination rule is programmed in the system and generates geminates from singletons by using a simple durational control method. It generates geminates by lengthening the duration of the consonant part of the syllables following the gemination mark. Two types of rules were prepared for two groups of consonants, continuant (voiced and unvoiced) and non-continuant (stops and glottalized) consonants. If a gemination mark (‘) is followed by syllable with voiced or unvoiced consonant then, the last three frames of the cepstral parameters (c[0]) of vowel is adjusted linearly and then 120 ms of frame 1, 2 and 3 of second syllable is added. Then the second syllable is connected after frame 4. Totally 90 ms of cepstral parameters is added. Otherwise, if, a gemination mark (‘) is followed by syllable with glottal or non-glottal consonant then, the last three frames of the cepstral parameters (c[0]) of vowel is adjusted linearly and then 100 ms of silence is added. Finally the second syllable is directly connected.

Since Amharic orthography does not use gemination mark, in our study we used our own gemination mark and a manual gemination insertion mechanism for input texts. Although some scholars make use of two dots (¨ which is proposed in UNICODE 5.1 version as 135F) over a consonant to show gemination, so far there is no software which supports this mark.

3.3.2 Sixth order syllables rule

As mentioned earlier the sixth order syllables are very frequent and play a major role for proper pronunciation of words. The sixth order orthographic syllables, which do not have any vowel unit associated to it in the written form, may associate the helping vowel (epenthetic vowel /ix/, see the Apendix) in its spoken form (Dawkins, 1969). The sixth order syllables are ambiguous; they can stand for either a consonant in isolation or a consonant with the short vowel. In our study we prepared a simple algorithm to control the sixth order syllables duration. The algorithm to model the sixth order syllables duration uses the following rules:

1. The sixth order syllables at the beginning of word are always voweled (see /sxix/ in fig 5).
2. The sixth order syllables at the end of a word are unvoweled (without vowel) but, if it is geminated, it becomes voweled.
3. The sixth order syllables in the middle of words are always unvoweled (see /f/ in fig.5). But, if there is a cluster of three or more consonants, it is broken up by inserting helping vowel /ix/.

The following figures shows sample words synthesized by our system by applying the prosodic rules. Fig.5 and fig.7 shows the waveform and duration of words synthesized by applying both the gemination and sixth order syllables rules. Fig.4 and fig.6 shows the waveform of original words just for comparison purpose only. The two synthesized words are comparative words which differ only by presence or absence of gemination. In the first word /sxixfta/ ሽፍታ meaning ‘bandit’, the sixth order syllable /f/ is unvoweled (see fig.5). However, in the second word /sxixffixta/ ሽፍታ meaning ‘rash’, the sixth order syllable /f/ is voweled and longer /ffix/ (see fig.7) because it is geminated. In our previous study, Tadesse and Takara (2006), we observed that vowels are needed for singletons to be pronounced as geminates.

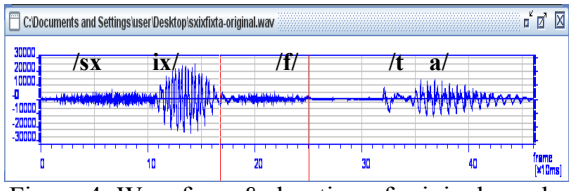

Figure 4: Waveform & duration of original word ሽፍታ/sxixfta/, meaning 'bandit'

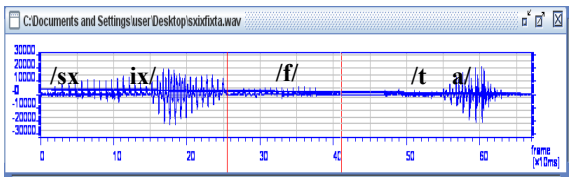

Figure 5: Waveform & duration of synthesized word ሽፍታ/sxixfta/, meaning 'bandit'

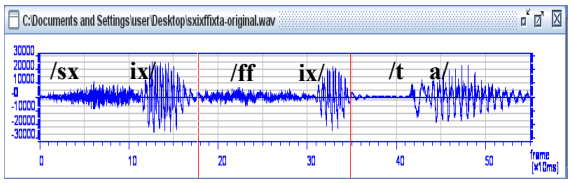

Figure 6: Waveform & duration of original word ሽፍታ/sxixffixta/, meaning 'rash'

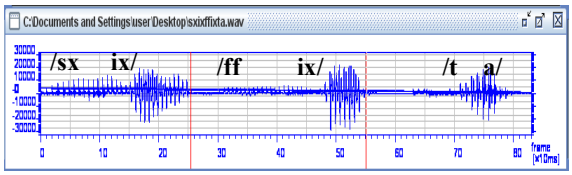

Figure 7: Waveform & duration of synthesized word ሽፍታ/sxixffixta/, meaning 'rash'

3.3.3 Syllables connection rules

For syllables connections, we prepared four types of syllables connection-rules based on the type of consonants. The syllable units which are represented by cepstrum parameters and stored in the database are connected based on the types of consonants joining with vowels. The connections are implemented either by smoothing or interpolating the cepstral coefficients, F0 and amplitude at the boundary. Generally, we drive two types of syllabic joining patterns. The first pattern is smoothly continuous linkage where the pitch, amplitude and spectral assimilation occur at the boundary. This pattern occurs when the boundary phonemes of joining syllables are unvoiced. Another joining pattern is interpolation, this pattern occurs when one or both of the boundary phonemes of joining syllables is voiced. If the boundary phonemes are plosive or glottal stop then the pre-plosive or glottal stop closure pauses with 40ms in length is inserted between them.

3.3.4 Intonation

The intonation for a sentence is implemented by applying a simple declination line in the log frequency domain adopted from similar study for Japanese TTS system by Takara and Jun (1988). Fig.8 shows the intonation rule. The time ti is the initial point of syllable, and the initial value of F0 (circle mark) is calculated from this value. This is a simple linear line, which intends to experiment the very first step rule of intonation of Amharic. In this study, we simply analyzed some sample sentences and take the average slope = -0.0011. But as a future work, the sentence prosody should be studied more.

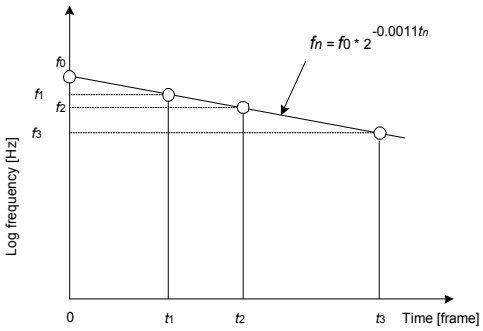

Figure 8: Intonation rule

4 Evaluation and Discussion

In order to evaluate the intelligibility and naturalness of our system, we performed two types of listening tests. The first listening test was performed to evaluate the intelligibility of words synthesized by the system and the second listening test was to evaluate the naturalness of synthesized sentences. The listening tests were used to evaluate the effectiveness of the prosodic rule employed in our system.

4.1 Recordings

For both listening tests the recording was done in a soundproof room, with a digital audio tape recorder (SONY DAT) and SONY ECM-44S Electrets Condenser Microphone. Sampling rate of DAT was set at 48kHz then from DAT the recorded data were transferred to a PC via a digital audio interface (A/D, D/A) converter. Finally, the data was converted from stereo to mono; down sampled to 10 kHz and the amplitude was normalized using Cool Edit program. All recording was done by male native speaker of the language who is not included in the listening tests.

4.2 Speech Materials

The stimuli for the first listening test were consisted of 200 words which were selected from Amharic-English dictionary. The selected words are commonly and frequently used words in the day-to-day activities adopted from (Yasumoto and Honda, 1978). Among the 200 words we selected, 80 words (40% of words) contain one or more geminated syllables and 75% of the words contain sixth order syllables. This shows that how geminates and sixth order syllables are very frequent.

Then, using these words, two types of synthesized speech data were prepared: Analysis/synthesis sounds and rule-based synthesized sounds using AmhTTS system. The original speech sounds were also added in the test for comparison purpose.

For the second listening test we used five sentences which contains words with either geminated syllables or sixth order syllables or both. The sentences were selected from Amharic grammar book, Baye (2008) which are used as an example. We prepared three kinds of speech data: original sentences, analysis/synthesis sentences, and synthesized sentences by our system by applying prosodic rules. In total we prepared 15 speech sounds.

4.3 Methods

Both listening tests were conducted by four Ethiopian adults who are native speakers of the language (2 female and 2 male). All listeners are 20-35 years old in age, born and raised in the capital city of Ethiopia. For both listening tests we prepared listening test programs and a brief introduction was given before the listening test.

In the first listening test, each sound was played once in 4 second interval and the listeners write the corresponding Amharic scripts to the word they heard on the given answer sheet.

In the second listening test, for each listener, we played all 15 sentences together and randomly. And each subject listens to 15 sentences and gives their judgment score using the listening test program by giving a measure of quality as follows: (5 – Excellent, 4 - Good, 3 - Fair, 2 - Poor, 1 – Bad). They evaluated the system by considering the naturalness aspect. Each listener did the listening test fifteen times and we took the last ten results considering the first five tests as training.

4.4 Results and discussion

After collecting all listeners' response, we calculated the average values and we found the following results.

In the first listening test, the average correct-rate for original and analysis-synthesis sounds were 100% and that of rule-based synthesized sounds was 98%. We found the synthesized words to be very intelligible.

In the second listening test the average Mean opinion score (MOS) for synthesized sentences were 3.2 and that of original and analysis/synthesis sentences were 5.0 and 4.7 respectively. The result showed that the prosodic control method employed in our system is effective and produced fairly good prosody. However, the durational modeling only may not be enough to properly generate natural sound. Appropriate syllable connections rules and proper intonation modeling are also important. Therefore studying typical intonation contour by modeling word level prosody and improving syllables connection rules by using quality speech units is necessary for synthesizing high quality speech.

5 Conclusions and future works

We have presented the development of a syllabic based AmhTTS system capable of synthesizing intelligible speech with fairly good prosody. We have shown that syllables produce reasonably natural quality speech and durational modeling is very crucial for naturalness. However the system still lacks naturalness and needs automatic gemination assignment mechanisms for better durational modeling.

Therefore, as a future work, we will mainly focus on improving the naturalness of the synthesizer. We are planning to improve the duration model using the data obtained from the annotated speech corpus, properly model the co-articulation effect of geminates and to study the typical intonation contour. We are also planning to integrate a morphological analyzer for automatic gemination assignment and sophisticated generation of prosodic parameters.

References

Atelach Alemu, Lars Asker and Mesfin Getachew. 2003. *Natural Language Processing For Amharic: Overview And Suggestions for a Way Forward*, Proc. 10th Conference 'Traitement Automatique Des Languages Naturalles', pp. 173-182, Vol.2, Batz-Sur-Mer, France.

Sebsibe H/Mariam, S P Kishore, Alan W Black, Rohit Kumar, and Rajeev Sangal. 2004. *Unit Selection Voice for Amharic Using Festvox*, 5th ISCA Speech Synthesis Workshop, Pittsburgh.

M.L Bender, J.D.Bowen, R.L. Cooper and C.A. Ferguson. 1976. *Language in Ethiopia*, London, Oxford University Press.

M. Lionel Bender, Hailu Fulass. 1978. *Amharic Verb Morphology: A Generative Approach*, Carbondale.

Tadesse Anberbir and Tomio Takara. 2006. *Amharic Speech Synthesis Using Cepstral Method with Stress Generation Rule*, INTERSPEECH 2006 ICSLP, Pittsburgh, Pennsylvania, pp. 1340-1343.

T. Takara and T. Kochi. 2000. *General speech synthesis system for Japanese Ryukyu dialect*, Proc. of the 7th WestPRAC, pp. 173-176.

Baye Yimam. 2008. **የአማርኛ ሰዋሰው** (“*Amharic Grammer”)*, Addis Ababa. (in Amaharic).

C.H DAWKINS. 1969. *The Fundamentals of Amharic*, Bible Based Books, SIM Publishing, Addis Ababa, Ethiopia, pp.5-7.

S. Furui, *Digital Speech Processing, Synthesis, and Recognitio*n, Second Edition, Marcel Dekker, Inc., 2001, pp. 266-270.

S. Imai. 1980. Log *Magnitude Approximation (LMA) filter*, Trans. of IECE Japan, J63-A, 12, PP. 886-893. (in Japanese).

Takara, Tomio and Oshiro, Jun. 1988. *Continuous Speech Synthesis by Rule of Ryukyu Dialect*, Trans. IEEE of Japan, Vol. 108-C, No. 10, pp. 773-780. (in Japanese)

B. Yasumoto and M. Honda. 1978. *Birth Of Japanses*, pp.352-358, Taishukun-Shoten. (in Japanese).

Appendix

Amharic Phonetic List, IPA Equivalence and its ASCII Transliteration Table. (Mainly adopted from (Sebsibe, 2004; Baye, 2008)

IPA Equivalent	Transcription	Amharic Scripts
Consonants		
[p]	[p]	ፐ
[t]	[t]	ተ
[k]	[k]	ከ
[ʔ]	[ax]	ዐ
[b]	[b]	በ
[d]	[d]	ድ
[g]	[g]	ግ
[p’]	[px]	ጵ
[t’]	[tx]	ጥ
[c’]	[cx]	ፂ
[q]	[q]	ቅ
[f]	[f]	ፍ
[s]	[s]	ስ
[ʃ]	[sh]	ሽ
[h]	[h]	ሀ
[s’]	[sx]	ጽ
[tʃ]	[c]	ች
[g’]	[j]	ጅ
[m]	[m]	ም
[n]	[n]	ን
[n’]	[nx]	ኝ
[l]	[l]	ል
[r]	[r]	ር
[j]	[y]	ይ
[w]	[w]	ው
[v]	[v]	ቭ
[z]	[z]	ዝ
[z’]	[zx]	ዠ
Vowels		
[ə]	[e]	ኧ
[ʊ]	[u]	ኡ
[ɪ]	[ii]	ኢ
[a]	[a]	አ
[e]	[ie]	ኤ
[i]	[ix]	እ
[o]	[o]	ኦ