Shared Syllables for Amharic Tigrigna Text to Speech Synthesis

Lemlem Hagos[1(✉)], Million Meshesha[1], Solomon Atnafu[2], and Solomon Teferra[1]

[1] School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia
{lemlem.hagos, million.meshesha, solomon.tefera}@aau.edu.et
[2] Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia
solomon.atnafu@aau.edu.et

Abstract. In this study, an experiment is conducted to explore and exploit shared Amharic and Tigrigna syllables in the development of Amharic Tigrigna bilingual text to speech synthesizer. Both Amharic and Tigrigna are under resourced languages, yet these two languages share the Geez writing system with large portion of phone sets and syllables. This study therefore shows the possibility of constructing Amharic-Tigrigna bilingual text to speech synthesizer based on the shared syllables to optimize linguistic resources. The dataset for training and testing is composed of consonant-vowel syllables in both languages. Festival speech synthesis framework is used for the experiment. The result shows mean opinion score of 3.09 and 2.08 for intelligibility and naturalness, respectively. Epenthesis vowel insertion and possibility geminates which are not predictable from the text at surface level in both languages greatly affect naturalness of the synthetic speech. Another factor that affects the naturalness is the fact that we used an already existing multilingual speech synthesis framework that has foreign accent. Even though the naturalness is below average because of the aforementioned reasons, the possibility of exploiting shared features to develop multilingual speech synthesis for under resourced languages is encouraging. We have learned that to enhance the performance of the bilingual synthesizer, there is a need to integrate language specific features.

Keywords: Bilingual text to speech · Festival · Syllable · Amharic-Tigrigna

1 Introduction

With the advancement in speech synthesis technology, researchers are aiming to achieve more natural sounding and intelligible speech output. Text to speech synthesis enables computers to convert arbitrary text into audible speech [1]. Text to speech synthesis undergoes the process of text analysis and speech generation [2].

The text analysis is responsible for determining the underlying structure of the sentence and the phonemic composition of each word. This is because of the fact that strings of phonemes form larger units such as syllables; which in turn form words,

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2022
Published by Springer Nature Switzerland AG 2022. All Rights Reserved
M. L. Berihun (Ed.): ICAST 2021, LNICST 411, pp. 550–558, 2022.
https://doi.org/10.1007/978-3-030-93709-6_37

constituting phrases and sentences. These structures need to be indicated in the underlying representations for an utterance, because aspects of how a sentence is pronounced depends on the locations of these types of boundaries showing pronunciation of each word, syntactic structure for the sentence and semantic focus to resolve ambiguity [1].

Speech generation part of text to speech synthesizer transforms the abstract linguistic representation into speech waveform. It is responsible for phonetic realization of each phoneme [1]. The speech synthesis part is also concerned with the selection and concatenation of appropriate concatenative speech units given the phoneme string as well as a speech waveform [2].

Nowadays, text to speech research has been pursued for the various languages in the world, such as English, most European and Asian languages [3, 4]. However, text to speech synthesis is at its infancy stage for under resourced languages including Ethiopian Languages. Having limited resources both data and tools, the aim of this study is to explore the possibility of designing a bilingual text to speech synthesizer for Amharic and Tigrigna, which are the two leading Semitic languages in Ethiopia as a medium of communication.

In the following sections, we present related works, features of Amharic and Tigrigna languages, experimentation, discussion of synthesis results and concluding remarks.

2 Related Research Works

There are studies that attempt to construct a model for Amharic as well as Tigrigna text to speech synthesis. Most of the research works are for Amharic and few are for Tigrigna.

Laine [5] initiated the first text to speech synthesis for Amharic language using diphone based concatenative technique in 1999. Tools used were Pascal and MATLAB programming languages, and the user acceptance evaluation is reported as good. Furthermore, Laine proposed the need for smoothing techniques to improve the performance of the synthesizer.

Henock [6] then applied TD-PSOLA technique for smoothing the concatenative speech units for Amharic. Tools and techniques used include PRAAT for spectrographic analysis and Delphi as well as MATLAB for developing Amharic synthesizer. It is reported that the evaluation of the models using ORT (Open Rhyme Test), and MOS (Mean Opinion Score), is promising and suggested consideration of prosodic elements for further investigation.

Tesfay [7] attempted the first text to speech for Tigrigna language using diphone based concatenative approach in 2004. MATLAB is used for implementation. The performance of the synthesizer is reported as MOS of 3.05. Inclusions of acronym converter to the text processing module and prosody control are issues noted by Tesfay as a way forward to enhance the performance of Tigrigna text to speech synthesis.

Nadew [8] applied formant-based speech synthesis for Amharic vowels using MATLAB. The focus was on vowels since vowels play a big role in changing pronunciation of a word in different contexts. Result indicated intelligibility of 88.85% for isolated vowels. Nadew recommended refinement of the work including consonant consideration, and preparation of standard speech corpus for training and testing.

Bereket [9] modeled an HMM based speech synthesizer with the objective of developing unlimited domain speech synthesizer for Amharic language that can generate a natural sounding and intelligible synthetic speech with less resource requirement. Out of 11,670 sentences 500 of them were used to train the HTS, and 20 sentences were used for testing. The performance result shows MOS of 4.12 and 3.6 for intelligibility and naturalness, respectively. As a future work, Bereket recommended inclusion of prosodic information to identify dialects and word meanings.

Experiment by Alula [10] explored the possibility of including non-standard words (NSWs) in Amharic text to speech synthesis. Alula used diphone based concatenative synthesis and RELP (Residual Excited Linear Predictive) coding. Performance of the synthesizer shows MOS of 3.0 for intelligibility and MOS of 2.83 for naturalness. Alula suggested the need for consideration of all types of NSWs and incorporation of part of speech (POS) tagged corpus for prosody control as a future work.

Each of the research attempts so far are focused on designing mono-lingual text to speech for Amharic and Tigrigna languages. However, with the existence of more than 80 under resourced languages in Ethiopia there is a need to study and exploit the common characteristics of related language family so as to develop a multilingual text to speech synthesis. To begin with, in this paper Amharic-Tigrigna bilingual text to speech synthesizer is reported based on their consonant-vowel (CV) syllable overlap, as per the way forward presented in [11]. The research output can be enhanced by applying transfer learning from resourced languages [12] and [13].

3 Features of Amharic and Tigrigna Languages

Ethiopian Semitic language is a sub family of south-Semitic which in turn is a sub family of West Semitic language under the Semitic language of the Afro-Asiatic super family. It includes Geez, Tigrigna, Tigre, Amharic and Argoba [14]. Amharic and Tigrigna are the second and the third most spoken Semitic languages in the world, next to Arabic [15].

Tigrigna has 29 consonantal phonemes and seven vowels [16]. According to Girmay [16], the plosive labiovelars, ጕ [gw], ኩ [kw], ቍ [k'w], as well as the fricative labiovelars, ኹ [xw] and ቖ [xw] are derivable from their respective core consonantal segments. Furthermore, Girmay notes that the fricative velars, ኻ [x] and ኾ [x'], are allophones of ክ [k] and ቅ [k'] respectively. Thus, these are not included in the consonantal chart of Tigrigna. According to Tsehaye [17] and Daniel [18], however, aforementioned derivable and allophone segments as well as the phoneme ቭ [V] are included in the consonant chart of Tigrigna. As a result, the number of consonants in

Tigrigna would be 37, which is composed of 32 core consonants and 5 labialized composite segments. The seven vowels are attached to the core consonants to create Tigrigna characters, which are CV syllables. Each of the five labialized segments provide five variants of characters.

Similarly, Amharic contains seven vowels and debatable number of consonants. Baye [19] argues that Amharic has 30 consonants, whereas Mulugeta [20] reduces them to 21 underlying consonants and 6 derivable palatal consonants. For the purpose of our experiment we took Baye's recommendation, with more consonants for Amharic. The character set in Amharic is made up of the 30 consonants by the seven vowels matrix together with 13 incomplete segments which are composed of a consonant followed by short w and the vowel a, as in ሟ/mwa/, ሷ/swa/ and ሯ/rwa/.

Amharic and Tigrigna share about 30 consonants in addition to the seven vowels. The majority of the character set in both Amharic and Tigrigna is composed of CV syllable. Furthermore, syllable structure of Amharic and Tigrigna shows some overlap. According to Baye [19], syllable structure of Amharic is V, VC, VCC, CV, CVC, and CVCC, where V stands for vowel, and C for consonant. On the other hand, Tsehaye [17] and Tesfay [21] noted that syllabic structure of Tigrigna is CV and CVC. Hence, both Amharic and Tigrigna share the CV and CVC syllable structure. The most common shared syllable across Amharic and Tigrigna is CV as both use character set that translate into CV syllable.

4 Experimentation

This study explores the syllable overlap between Amharic and Tigrigna languages. Accordingly, an experiment is conducted towards designing a bilingual text to speech synthesis based on the most frequent CV syllables selected from both Amharic and Tigrigna text. To undertake the experiment, festival text to speech synthesis framework is used since it is an open source multilingual text to speech synthesis framework.

4.1 Dataset Preparation

A dataset of 2000 sentences was purposefully selected from newspapers covering a wide range of issues including socioeconomic and politics among others. The prepared dataset is composed of 1000 sentences from each of Amharic and Tigrigna texts in order to gain good coverage of the character sets in both languages. An increase in the number of sentences used for the purpose of training and testing provides better coverage of the CV dataset.

Phonetic transcription is done using a mapping table that translates each Ethiopic character into its phonetic equivalent, which is usually a consonant-vowel (CV) combination. There are instances however where a consonant is transcribed into the same consonant followed by a short vowel or epenthesis vowel. It is observed that Amharic makes use of less epenthesis than Tigrigna. Accordingly, we tried to transcribe the sixth order character in Amharic as a consonant, while that in Tigrigna is transcribed with the epenthesis vowel following the consonant.

After the transcription of the text, there was insertion of epenthesis vowel in times where there is a cluster of more than two consonants in Amharic, as well as deletion of inserted epenthesis for Tigrigna, especially at word final position.

4.2 Identifying Syllable Overlap

Once the dataset is prepared, we then explore syllable overlap that exists between Amharic and Tigrigna languages. Consonant-vowel (CV) syllables are then counted using python programming language. Tigrigna character set is composed of 32 consonants by 7vowels plus 5 derivable labialized segments each having 5 variants, which makes up 249 characters. We found out that the Tigrigna dataset is composed of 195 distinct characters. Thus, the Tigrigna character set coverage is 78.63%.

Similarly, Amharic character set consists of 30 consonants by 7 vowels plus 13 incomplete characters such as ኳ/kwa/ which are derivable from core consonants followed by short w and then the vowel a. This makes the required number of characters in Amharic to 223. The prepared Amharic dataset is composed of 211 unique characters. The coverage of the character set in Amharic dataset is therefore, 94.61%. Even though the dataset is composed of 1000 sentences for each Amharic and Tigrigna languages, the sentence length in these languages is different as the texts are collected from different sources with different authors. Hence, Amharic sentences are longer compared to Tigrigna sentences. As a result, we found more character coverage in Amharic dataset than in Tigrigna.

Amharic and Tigrigna text is analyzed for the consonant-vowel syllable overlap and found to be 70.51%. This shows that the CV level overlap is a little bit below the actual overlap because of lower coverage of the characters for Tigrigna dataset, which in turn is attributed to the shorter sentences in Tigrigna text.

The most frequent and shared syllables are identified and words that contain these shared syllables are used as an input for the synthesizer. Words that contain the most frequent CV syllables are selected from the aforementioned dataset and used as an input to the synthesizer.

Syllables are sensible to the human ear when their context in a word is understood. Words are also more sensible when embedded in a sentence. In both Amharic and Tigrigna, a syllable can assume a word initial, word medial or word final position. These are the contexts that are considered in this experiment. Accordingly, the words shown in Table 1 below are selected with the intention of including these contexts.

In order for the Festival speech synthesis system to recognize the Amharic and Tigrigna text input, SERA algorithm is applied so as to convert the Ethiopic orthography into its equivalent ASCII representation.

Table 1. Data set for experiment

syllable	Tigrigna			Amharic		
	Word Initial	Word Medial	Word Final	Word Initial	Word Medial	Word Final
ን n/nɨ	ንምህዞ nimihzo	መንጎ məngo	ተግባርን təgbarrin	ንግግር nigigir	ድንቅ dink’	ስኬትን siketin
ብ b/bɨ	ብሞያ bimoja	መብርሂ məbrɨhi	ሰብ səb	ብርሃን bɨrhan	ኣዳብራ ʔadabra	ሃሳብ Hasab
ት t/tɨ	ትግራይ tɨgraj	ፍትሒ fɨtɨhi	ከባቢታት kəbabitat	ትጋቱ tɨgatu	ማትረፍ matrəf	ሽልማት ʃɨllɨmat
ም m/m(ɨ)	ምኽንያት mɨxnɨjat	ተምቤን təmben	ኣክሱም ʔaksum	ምህረት mihrət	በማምረት bəmamrət	ሰላም səlam
በ/bə	በለ bələ	ማሕበራዊ maħbərawi	ግንዛበ gɨnizabə	በዘርፉ bəzərfu	ለበርካታ ləbərkata	ያልታሰበ jaltasəbə
ለ/lə	ለኣኩላ ləʔakula	መለለይ mələləj	ጉጅለ gujɨlə	ለማዘጋጀት ləmazəgaj ət	ባለሙያ baləmuja	በተሻለ bətəʃalə
ተ/tə	ተወሊዱ təwəlidu	እተን ʔitən	ክልተ kɨlɨtə	ተከታታይ təkətatj	በተገቢው bətəgəbiw	እየከፈተ ʔɨjəkəfətə
መ/mə	መንነት mənɨnət	ዝተመስረተ zɨtəməsrətə	ደጋጊመ dəgagimə	መዘጋጀት məzəgajət	በመፍጠር bəməft’ər	እየታተመ ʔɨjətatəmə
ረ/rə	ረብሓ rəbħa	ኣረምኩ ʔarəmku	ዝነበረ zɨnəbərə	ረዳት rəddat	ደረጃ dərəjja	ከጀመረ kəjəmmərə
ዝ/zə	ዝምደብ zɨmidəb	ብዝርኣይ bizɨrʔaj	እትጉዓዝ ʔɨtɨguʕaz	ዝርዝር zɨrzɨr	አብዝተው ʔaβzɨtəw	አልማዝ ʔalmaz
የ/jə	የራኢ jəraʔi	ኣካየዱ ʔakajedu	ሰንየ sənijə	የያዘው jəjazəw	አየለ ʔajələ	ተለየ tələjə

5 Discussions of Synthesis Results

The extent to which each syllable is clearly heard in the context of the word it embeds is judged by bilingual speakers. The evaluation is based on the five scale mean opinion score (MOS). For evaluation, words that contain the syllable to be evaluated in three different contexts (word initial, word medial, and word final) for the two languages are considered. Thus, the evaluator first looks at each of the six words for every one of the eleven selected syllables. Then he/she is presented with the synthetic speech and asked to specify the level of comprehensibility of each syllable with the scale ranging from poor to excellent. The same 66 words are used to evaluate naturalness, the extent to which the synthetic speech that contains the eleven syllables are pleasant or tolerable or annoying to the human ear.

The evaluation result shows that the synthesizer registers a mean opinion score of 3.09 for intelligibility and 2.08 for naturalness. The intelligibility of the synthesizer is at acceptable level of producing understandable synthetic speech. The decrease in naturalness of the synthetic speech is on the other hand related to the fact that some of the

characters with unique sound, found in both Amharic and Tigrigna such as families of ቀ, ቐ, ሐ, ጠ, and ዕ, are missing from the festival speech synthesis engine. Even though few characters such as members of �‘ and ኧ are allophones of ቀ and ከ respectively [16], they have their own unique signal spectrograph and there is a need for investigating the signal level relationship among allophones to identify the base signal and the factor to generate the derivable signals. Furthermore, epenthesis vowel and geminates greatly affect naturalness of the synthetic speech as they are expecting the integration of a well-crafted rules during post processing.

Furthermore, variation in intelligibility and naturalness of the same syllable is observed according to the context of the syllable which appears at a word initial, word medial or word final position, as shown in Table 2. Analysis of the context of the syllables, shows that, the average intelligibility and naturalness get the best result at the word initial position of the syllable. Average result is low at the word medial context. This is because of the influence of the surrounding syllables which induce co-articulation effect.

Table 2. Mean opinion score

	Word initial	Word medial	Word final	Average
Intelligibility	3.23	3	3.05	3.09
Naturalness	2.18	2	2.05	2.08

In addition, since the syllables are synthesized on a generic multilingual synthesizer, festival, the naturalness of the synthesized speech is greatly affected as compared to intelligibility of the synthesis result. This is because the generic synthesizer lacks to map the inherent sounds represented in the syllables. This requires integration of the unique Ethiopic sounds observed in Amharic and Tigrigna speech and we are working towards this direction as our future work.

6 Concluding Remarks

In this research, we have explored the consonant vowel syllables that are common for Amharic and Tigrigna to build a bilingual text to speech synthesizer. Amharic and Tigrigna share the same Geez writing system as well as a large portion of their phone sets and syllables. We selected purposefully 2000 sentences of Amharic and Tigrigna texts from newspapers and analyzed consonant vowel syllables across the two languages.

Once the CV overlap between the two languages is analyzed, the most frequent and common consonant vowel syllables are selected to implement a bilingual text to speech synthesizer for these languages. Words containing frequent syllables are selected from the prepared dataset so that the syllables assume three different contexts according to their position in a word, word initial, word medial and word final. Festival speech synthesis framework is used for the experiment.

The performance of the bilingual synthesizer is evaluated for its intelligibility and naturalness by bilingual speakers of the languages. The average performance of the synthesizer, taking into account all three possible contexts of the syllables, is MOS of 3.09 and 2.08 for intelligibility and naturalness, respectively. The result shows that there is a variation of performance in relation to the position of the syllable, where better performance in both intelligibility and naturalness is attained at word initial context; and the least performance at the word medial context. The decrease in intelligibility and naturalness of the word medial syllables is attributed to the effect of the surrounding syllables.

In general, the evaluation result indicates that intelligibility of the synthetic speech is at acceptable level and naturalness needs to be improved further by investigating signal level mappings for the unique sounds in Amharic and Tigrigna, which is our next research direction.

References

1. Taylor, P.: Text to Speech Synthesis. Cambridge University Press, New York (2009)
2. Strout, R., Olive, J.: Text to speech synthesis. In: Vijay, D.B.W., Madisetti, K. (ed.) Digital Signal Processing Handbook, pp. 976–986. CRC Press, Lonndon (1999)
3. Lemmetty, S.: Review of Speech Synthesis Technologies. Helsinki University of Technology, Helsinki (1999)
4. Sagisak, Y.: Spoken output technologies. In: Survey of the State of the Art in Human Language Technology, pp. 165–197. Cambidge University Press, Cambidge (1997)
5. Laine, B.: Text to Speech Synthesis for Amharic Language. Addis Ababa University, Addis Ababa (1999)
6. Henock, L.: Concatenative Text to Speech Synthesis for Amharic language. Addis Ababa Univerity, Addis Ababa (2003)
7. Tesfay, Y.: Diphone Based Text to Speech Synthesis for Tigrigna language. Addis Ababa University, Addis Ababa (2004)
8. Nadew, T.: Formant Based Synthesis for Amharic Vowels. Addis Ababa University, Addis Ababa (2008)
9. Bereket, K.: Developing a Speech Synthesizer for Amharic using Hidden Markov Model. Addis Ababa University, Addis Ababa (2008)
10. Alula, T.: A Generalized Approach to Amharic Text to Speech (TTS) Synthesis System. Addis Ababa University, Addis Ababa (2010)
11. Lemlem, H., Million, M.: Text to speech synthesis for ethiopian semitic languages: issues and the way forward. In: 12th IEEE Africon International Conference, Addis Ababa (2015)
12. Chen, Y.-J., Tu, T., Yeh, C.-C., Lee, H.-Y.: End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning. arXiv e-print arXiv:1904.06508v2 (2019)
13. Lee, Y., Shon, S., Kim, T.: Learning pronunciation from a foreign language in speech synthesis networks. arXiv e-prints: arXiv.1811.09364v4 (2020)
14. Bender, L., Hailu, F.: Amharic Verb Morphology: A Generative Approach, Michiga Michiga State University (1978)
15. Ethnologue Homepage. https://www.ethnologue.com. Accessed 15 June 2017
16. Girmay, B.: The Phonology of Tigrigna: Generative Approach. Addis Ababa University, Addis Ababa (1983)

17. Tsehaye, T.: Reference Grammar of Tigrigna. Georgetown University, Washignton DC (1979)
18. Daniel, T.: Modern Tigrigna Grammar. Biranna Press, Addis Ababa (2008)
19. Baye, Y.: Amharic Grammar. Elleni Press, Addis Ababa (2007)
20. Mulugeta, S.: The Syllable Structure and Syllabification in Amharic. Norwegian University of Science and Technology, Oslo (2001)
21. Tesfay, T.: A Modern Grammar of Tigrigna, Tipografia U. Detti, Rom (2002)