Design and Development of a Text-to-Speech Synthesizer for Afan Oromo

Tamrat Delessa Chala1,2 · Ashenafi Chalchisa Guta1 · Muluken Hussen Asebel1

Received: 24 August 2020 / Accepted: 2 July 2022 © The Author(s) 2022

Abstract
Speech is one of the natural ways of communication between humans, later extended as a means for human–computer interaction. It helps visually impaired people to read electronic texts and is used in information retrieval and language education. This paper proposed the development of a text-to-speech synthesizer for Afan Oromo (Oromo Language), using unit selection speech synthesizer approaches. Although several works have been conducted in the area of text-to-speech synthesis for technologically favored languages for many years, every language has its own unique features. So, speech synthesizer systems developed for one language cannot be used for another language, because the structures of one language are not presumably representative of others. It is clear that each program is based on the system corresponding to the phonetic rules of a certain language. Besides, the existing text-to-speech synthesizer for Afan Oromo was reviewed in this study and the result of developed prototype results are showing promising, however, still, their performance needs a lot of improvement in terms of intelligibility and naturalness using novel approaches and quality of corpus. Therefore, this research was initiated to develop the possibility of developing a prototype text-to-speech synthesizer to improve the performance of the text-to-speech synthesizer. In this study, Afan Oromo corpus was collected from genuine sources and prepared speech datasets both text and audio in collaboration with Afan Oromo experts. The performance of the synthesizer was tested by proper users for its intelligibility and naturalness using Mean Opinion Scale (MOS). The obtained result of naturalness of the prototype is 4.44 (very good) out of 5, which indicated that the result obtained is encouraging and better performance than the existing TTS of Afan Oromo in terms of intelligibility and naturalness. But the result scored in terms of intelligibility still needs further work. The main challenge is Afan Oromo has many dialects, so preparing a balanced text corpus from each dialect is very tough. Moreover, enhancement of the work is predicted to bring a reasonable level of intelligibility to the system.

Keywords Naturalness · Intelligibility · Speech synthesizer

Introduction
The text-to-speech synthesis (TTS) is the automatic conversion of unrestricted natural language sentences in text form to a spoken form that closely resembles the spoken form of the same text by a native speaker of the language. Additionally, text-to-speech is a type of speech synthesis application that is used to create a spoken sound version of the text in a computer document, such as a help file or web page. TTS can enable the reading of computer display information for visually impaired persons, or may simply be used to augment the reading of a text message [1].

Text-to-speech has a massive collection of applications, out of which the first and the most essential is the one used in reading systems for the visually impaired, where a system would read some texts from a written source or book and convert them into speech. These existing systems of course sounded very mechanical, but their adoption by visual impaired people was hardly surprising as the other option of reading Braille or having a real person do the reading for them were often not possible. Nowadays, quite sophisticated systems exist that
facilitate human computer interaction for the visual impaired, in which the TTS can help the user, navigate around windows system as a block diagram of a general TTS depicts in Fig. 1. The TTS can be distinguished as front-end which is the part of the system closer to the text input and back-end that is closer to the speech output [2].

Many researchers have been studying for centuries for artificial production of speech. The effort has transited from mechanical modeling of human speech production system to electrical speech synthesizers and now to different modern synthesis techniques of concatenating recorded speech with text analysis to obtain more natural sounding voice output [3]. Speech synthesizers have been developed for the language spoken in the developed countries for decades. However, there are few attempts for Ethiopia languages like Afan Oromo (Oromo language). Afan Oromo is a Cushitic language spoken by about 40 million people in Ethiopia (about 40% of the country’s population), Kenya, Somalia and Djibouti, and it is the third largest language in Africa after Arabic and Hausa. Afan Oromo (meaning Oromo language) is rated second widely used language among the indigenous languages of Africa [4]. While exploring the history of the Oromo nation, Gada [5] claimed that, Afan Oromo was the mother tongue of about 30 million Oromo people living in Ethiopia and its neighboring countries a couple of decades ago. For this language with numerous users, there have been only two attempts made, as far as the knowledge of the researchers is concerned, to develop a TTS for Afan Oromo. The first attempt to develop TTS for Afan Oromo was made by Morka [6] using the diphone speech units technique. However, in that study, prosodic features (features that look when we put the sound together in connected speech) were not considered and the performance of the developed prototype was highly affected. Similarly, Samson [7] conducted to develop a text-to-speech synthesizer for Afan Oromo using a diphone and triphone-based approach. Nevertheless, there are many shortages, including misrepresentation from discontinuities in the concatenation points. The developed system is time-consuming for training and recording speech, and the found result was a low performance in terms of intelligibility and naturalness. Furthermore, Diriba [8] conducted an automatic sentence parser for Afan Oromo based on a supervised learning technique. But, that study did not consider Afan Oromo TTS. On the other hand, Henock [9] and Tesfaye [10] developed a text-to-speech synthesizer for Amharic and Tigrigna (two other Ethiopian languages), respectively. But the nature and the structure of the languages are by far different from Afan Oromo.

Hence, to overcome these shortages, a new text-to-speech synthesizer for Afan Oromo is proposed using the unit selection synthesis approach. On the other hand, in this study, a finer speech dataset was used, and standard words were considered in collaboration with the Afan Oromo experts. The main objective of this study is to explore the possibility of developing a prototype text-to-speech synthesizer so that it can accept texts from a user and generate a natural language for Afan Oromo. More specifically, this research aims to:

- review related works to have a conceptual understanding of the area and identify the state-of-the-art in the text-to-speech synthesis;
- select better algorithm for building TTS system for Afan Oromo;
- design and implement a prototype text-to-speech synthesizer for Afan Oromo; and
- experiment and measure the performance of the system on the selected testing dataset.

In the last decades, the performance of speech processing system like speech synthesizer has improved dramatically, resulting in an increasingly widespread use of speech science and technology in real world scenario. Even though the TTS technology is growing from time to time to make things easy, people in developing countries like Ethiopia are unable to utilize it. This is because of the absence of speech synthesizers for local languages, limited access to information technology, lack of knowledge about foreign languages, and other economic issues to afford and use it.

It is not difficult to imagine that individuals with visual impairment need a system that helps them to understand what printed data they encounter says. If there are no text-to-speech synthesizing systems, people who are visually impaired cannot use computers to read and understand electronic materials like e-books, which makes them isolated and depressing. In addition to this, it is clear that a text-to-speech synthesizing system should be developed for every language because text-to-speech synthesizing has a lot of benefits like on websites, mobile apps, e-books, e-learning, conversational customer experience, media, transport experience, media, and robotics. There are fully functional (full-fledged) applications of text-to-speech synthesizer for languages like English, Spanish, Chinese and Hindi. Nevertheless, there is no such enhanced
text-to-speech synthesizing system implemented for local languages like Afan Oromo in Ethiopia.

Related Works

A text-to-speech (TTS) synthesizer is a computer-based system that can read text aloud automatically, regardless of whether the text is introduced by a computer input stream or a scanned input submitted to an optical character recognition (OCR) engine [2]. The speech synthesis procedure consists of two main phases, first converting the input text into a phonemic internal representation and then converting this internal representation into a waveform. The first step is text analysis and the second one is waveform synthesis. These two phases are usually called high-level and low-level synthesis, respectively [11]. A lot of papers are published on the text-to-speech synthesis for different languages around the world. Some papers are reviewed that the researchers thought relevant for the study.

Jayanthi et al. [12] proposed a new TTS which is based on a unit selection synthesis approach with syllable unit for Indian languages. The main purpose of the authors’ study was to enhance the synthesis quality and experimentation on the set of 1180 prompts, which focused on using larger or variable size units in synthesis. In that study, syllable segments of the continuous speech database were extracted and classified into begin, mid, and single units based on their position in words. As a result, the authors state that their method increased synthesis quality, and it reduced search space, improving synthesis effectiveness.

Borkar and Patil [13] conducted a study on the Text-to-Speech System for Konkani Language; they proposed a concatenation technique to develop this TTS system. To verify the validity of the proposed method, the researchers conducted experimentation on more than 1000 commonly used words in Konkani Language. The data sets were organized, and the wave files were recorded by students’ voices and around 3000 wave files consisting of vowels, characters, and half characters. In general, they stated that the concatenation method for the Konkani language was performed well because their system converted the complex word to speech easily.

In a separate study, Shreekanth et al. [14] conducted a unit selection-based Hindi TTS system using syllables as a basic unit. The study was implemented using.NET and MATLAB programming languages. To carry out that study, the researchers prepared a data set of about 1540 words from standard ‘Hindi to English dictionary’ and they recorded syllables with a sampling frequency of 16 kHz and represented them using 16 bits. Moreover, the researchers classified the syllable into three possible positions, namely the beginning of the word (start), between two syllables (middle), and at the end of the word (end).

Some researchers conducted TTS studies on different Ethiopian languages. For instance, Nadew [15] conducted a speech synthesis for Amharic, an Ethio-Semitic language, based on a formant based synthesis method, which is sometimes called synthesis by rule. The purpose of that study was to generate vowels from a given context, using the best selection of parameters from the decision tree. The researcher verified the validity of the proposed technique by conducting an experiment on five hundred (500) isolated Amharic words and digit strings that were recorded by twelve (12) different Amharic speakers. The author concluded that the proposed formant-based synthesis method provided high flexibility due to the potential of adjusting any of the acoustic parameters during run time. However, in that research, the researcher did not consider Amharic consonants. Hence, the author recommended that both vowels and consonants of Amharic should be considered to make the system fully functional.

In addition, Sebsibe [16] conducted a TTS system on Amharic by proposing unit selection voice for Amharic using festvox. Festivox is the tool that helps us in the creation and analysis of large speech corpora. In this study, the author developed a unit selection concatenative speech synthesizer by defining a transliteration scheme using ASCII standard characters. The system was designed based on the orthographic ordering of the scripts and the sound relationship of letters. The author also developed the phone set of the language with the phone’s corresponding features such as voicing, tongue position, tongue height, place of articulation, syllabification rules, and letter to sound rules into festvox. The researcher verified the validity of the proposed approach by conducting an experiment on 29,480 diphone corpus whose instances were made up of 801 unique diphones, and he tested the corpus consisting of a total of 12,724 syllables instances and 1317 unique syllables. The conclusion from that study was that the proposed system was suitable for the Amharic TTS system, and its performance levels ranged from excellent (5) to very poor (0), with a cumulative result of 2.9. Finally, the researcher recommended the optimal selection of corpus to produce better quality.

Tewodros [17], another researcher, carried out a text-to-speech synthesizer for Wolaytta—another Ethiopian language that belongs to the Omotic family. The approach the researcher adopted in this research was diphone concatenative synthesis based on residual LPC technique. As the researcher indicated, different experimentations were conducted to verify the validity of the proposed technique on 841 diphones dataset from 1156 hypothetical which were collected from different Wolaytta word corpus. The results
obtained from the experimentation showed a 78% performance average, 3.17 MOS score intelligibility and 2.77 MOS score naturalness. Finally, the researcher recommended that the phonetics and phonology of Wolaytta need to be considered to produce a better quality of the synthesizer.

All the reviewed studies reported that using the concatenative speech synthesis approaches is promising for TTS. However, the local studies conducted to develop TTS for Afan Oromo were only two pieces, although Afan Oromo is the most widely spoken language in Ethiopia [5]. From the reviewed studies the weakness of developed TTS for Afan Oromo was identified, and we have been trying to lay on the improvement of intelligibility and naturalness of TTS for Afan Oromo using unit selection approach and considering phonology of the language critically while keeping all the advantages of the earlier developed prototype. In addition, we cannot adopt TTS from other languages easily because it depends on the structure and nature of the languages. Therefore, the most effective methods and tools are used to implement a text-to-speech synthesizer for Afan Oromo.

### Phonology of Afan Oromo

Identifying the syllable structure of a language helps to understand the nature of the texts written in that specific language. The syllable is obligatory in any language to construct a sound and words. Therefore, this paper presents the phonology of Afan Oromo at glance to show how to train the machine for a good performance. Afan Oromo is one of the languages in Ethiopia that uses a customized Latin alphabet for its writing. The letters of the Afan Oromo alphabet are either vowels or consonants, and the relationship between letters and sounds in the language is one to one. The letters a, e, i, o, and u represent Afan Oromo vowels shown in the Table 1. All vowels are used basically in the same way throughout Afan Oromo dialects. Vowel length is contrastive or phonemic in the language. To show vowel length, these individual letters are repeated in writing: deemuu (go), nyaadhu (eat).

Consonants in Afan Oromo may come individually or in clusters, but must be connected to a vowel to form a syllable. Gemination is contrastive in Afan Oromo, and in writing consonants is repeated to indicate this.

Table 1 lists the consonants of Afan Oromo.

The syllable is a basic unit of speech studied on both the phonetic and phonological levels of analysis. Phonetically syllables are usually described as consisting of a center which has little or no obstruction to airflow and which sounds comparatively loud; before and after that center there is greater obstruction to airflow and/or less loud sound [19]. Laver [20] defines the phonological syllable as “a complex unit made up of nuclear and marginal elements”.

As in any language there are a set of rules that govern the correct construction of the various levels of language, i.e., syllables, words, and sentences. The syllabic structures of Afan Oromo are constrained by the following rules [21].

a. A word in Afan Oromo cannot start by two or more different consonants, and the syllable also shares this rule.

| Vowels (`Dubbachiiftuu`) [18] |
|-----------------------------|
| Front Mid Back |
| Close i/i/, ii/i:/ u/u/, uu/u:/ |
| Mid e/e/, ee /e:/ o/o/, oo/o:/ |
| Open a /A/ aa /a:/ |

| Consonants (`Sagaleewwan`) [18] |
|---------------------------------|
| Bilabial/ Labiodental Alveolar/ Retroflex Palato-alveolar/ Palatal Velar Glottal |
| Stops and affricates |
| Voiceless P t ch/tʃ/ k |
| Voiced B d j/dʒ/ g |
| Ejective ph/p'/ x/t'/ c/tʃ'/ |
| Implosive dh/d'/ |
| Fricatives |
| Voiceless f s sh/f/ H |
| Voiced v z |
| Nasals m n ny/n/ |
| Approximants w l y/y/ |
| Rhotic r |
Design and Implementation

In this study, both text and speech corpora were required for the development of a text-to-speech synthesizer for Afan Oromo. Text corpus was collected and preprocessed from Afan Oromo newspapers and different books written in Afan Oromo. Consequently, 1000 sentences were selected from various domains, including culture, politics, etc., by consulting Afan Oromo experts and marked phonetically. After the text corpus was prepared, the recording was made by native speakers reading sentence by sentence using Audacity, which is an open-source tool to help us for recording and edit sound. Finally, an audio corpus was preprocessed to remove noise using Praat, a freely available speech analysis tool. The text corpus is prepared as is shown in the following sample (Fig. 2):

To implement this research, Festival, and festivox tools were used for building speech synthesis systems for Afan Oromo. These tools were used to add voices and languages. The festival offered a general framework for building speech synthesis systems, as well as for including examples of various modules. As a whole, it offered full text to speech through a number APIs: from shell level, via a scheme command interpreter as C++ library, from Java, and an Emacs interface [22]. Even if the festival is one of the oldest running applications for text-to-speech on Linux, it is certainly a useful way of testing a text-to-speech-based application.

As presented earlier, to develop the prototype of this study, a festival speech synthesis framework was used, in which the process of converting text into speech has gone through a series of steps of designing the prompts, recording the prompts, auto label the prompts, build the utterance structure for recorded utterances, extract pitch mark and build LPC coefficients, build cluster unit based synthesizer from the utterances and finally testing.

All systems need to undergo rigorous testing before deployment, even if text-to-speech synthesizer testing is not as easy as other systems. It is widely agreed that a TTS system has two main goals in system test-making synthesized speech intelligible and natural. Intelligibility tests can be performed by word recognition tests or comprehension tests, where listeners are played a few words either in isolation or in a sentence and asked which words they heard. In the naturalness test, listeners are played some speech (phrase or sentence) and simply asked to rate what they hear. [23].

In this study, to evaluate the prototype, 6 words and 6 phrases were prepared. As shown in Tables 3 and 4, the selected Afan Oromo words and phrases are presented for one expert and two native speakers of Afan Oromo. In this

---

**Table 3** Performance measure TTS for Afan Oromo in terms of intelligibility

| Word     | P1 | P2 | P3 | Average |
|----------|----|----|----|---------|
| Otoo     | 5  | 5  | 5  | 5       |
| Sibila   | 2  | 2  | 1  | 1.67    |
| Ijoollee | 3  | 2  | 3  | 2.67    |
| Bishaan  | 4  | 3  | 3  | 3.33    |
| Hoodhu   | 1  | 1  | 1  | 1       |
| Seenaa   | 4  | 5  | 5  | 4.67    |
| Average  | 3.06 |    |    |         |

**Table 4** Performance measure TTS for Afan Oromo in terms of naturalness

| Phrases                  | P1 | P2 | P3 | Accuracy |
|--------------------------|----|----|----|----------|
| Alagaa gaafa             | 4  | 5  | 4  | 4.33     |
| Goondan wal-qabattee     | 4  | 4  | 5  | 4.33     |
| Yoo hubatani ilaalan     | 5  | 5  | 5  | 5        |
| Foon lafa jiru           | 5  | 5  | 5  | 5        |
| Otoo beektu huuba wajjin jette | 3  | 3  | 4  | 3.33     |
| Qalloon keessi sibila    | 4  | 5  | 5  | 4.67     |
| Average                  |    |    |    | 4.44     |

---

Fig. 2 Sample of Afan Oromo text corpus
Analysis and Discussion of Results

This study designed and developed a text-to-speech synthesizer and tested the intelligibility and naturalness of the system by one expert and two native speakers of the languages. In this testing, we tested at the word level for the system intelligibility and at the phrase level for the naturalness of the system. The testing results showed that the naturalness of the system is better than the intelligibility of the system, which scored 4.44 (very good). The accuracy of the synthesizer is perfect when the input is a phrase or sentences instead of words. It was observed that the system was trained by sentences because of this it performed well on phrases and sentences. On the other hand, the reason it performed less accurately at the word level was that Afan Oromo did not have its own parser.

Additional thought was that when the users were presented with synthetic speech for the first time, there is tough to recognize the principle of the synthetic speech. However, when the listeners get used to the synthetic speech, they become able to recognize the sound of the synthetic speech produced by the system. This is due to the adaptive nature of the human ear and the distance between synthetic speech and natural speech.

The overall result of this study to synthesizing text-to-speech synthesizer for Afan Oromo based on unit selection was good in terms of intelligibility. And, it was very good in terms of naturalness, i.e., it is promising and to the standard for increasing the training dataset as well as fine-tuning other parameters that affect the performance of the synthesizer.

Conclusion and Future Work

Nowadays, a text-to-speech-synthesizer enables students with physical disabilities and learning disabilities to recognize texts. It also ensures that texts communicate a writer’s intended meaning. This feature eliminates the stress of having to rely solely on visual cues to ensure written content is correct.

Afan Oromo is one of the Cushitic languages with millions of speakers in Ethiopia, especially in Oromia region. Therefore, to modernize Afan Oromo and to make it a language of science and technology, producing a text-to-speech synthesizer technology for the language is very important. This study proposed a unit selection synthesis approach to develop Afan Oromo text-to-speech synthesizer, the research involved collecting the sentences, preprocessing the sentences, recording the sentences by native speakers, preparing uniphone database, and developing the prototype of the tool using festival speech synthesis systems.

Consequently, the prototype was developed the intelligibility and naturalness of the system were tested by one expert of Afan Oromo and two native speakers. Using six words and six phrases test case the result obtained indicates the average accuracy of naturalness is 4.44/5 which is encouraging. The result scored for measuring the intelligibility showed that there is a need for further research to bring the intelligibility of the synthetic speech to read any word, phrases, and sentences of Afan Oromo.

Besides, by reviewing the relevant literature, we have identified some weak points of TTS for Afan Oromo, and thus, we proposed in this study a better approach, with the goal of further increasing the intelligibility and naturalness of TTS for Afan Oromo and compared to existing Afan Oromo TTS the new approach provides better in terms of intelligibility and naturalness.

Moreover, this research has developed a text-to-speech synthesizer for Afan Oromo; subsequently, the results obtained are promising. To increase the performance of the text-to-speech synthesizer for Afan Oromo, an integration with sophisticated machine learning algorithms is the future direction of the text-to-speech synthesizer for Afan Oromo. In addition, the enhancement of the work is expected to bring a reasonable level of intelligibility to the system.

Acknowledgements We thank Ministry of Communication and Information Technology for financial support.

Funding Open access funding provided by Széchenyi István University (SZE).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes...
were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Hunton JE. Blending information and communication technology with accounting research. Account Horiz. 2002;16:55–67.
2. Sasirekha D, Chandra E. Text to speech: a simple tutorial. Int J Soft Comput Eng. 2012;2:275–8.
3. Palo P, Laine UK. A review of articulatory speech synthesis. Master’s of thesis, University of Helsinki, 2006
4. Tolessa T. Early history of written Oromo language up to 1900. Sci Tech Arts Res J. 2013;1:76–80.
5. Melbaa Gadaa O. An introduction to the history of the Oromo people. Khartoum, Sudan, 1988
6. Morka M. Text- to- speech system for Afan Oromo. MSc, Addis Ababa University, 2001
7. Samson T. Concatenative text-to-speech system for Afan Oromo Language. MSc thesis Addis Ababa University, 2011
8. Diriba. An automatic sentence for Oromo language using supervised learning technique. M.Sc, Information Science Addis Ababa University 2002
9. H. L. Contaminative Text- to- speech system for Amharic Language. MSc, Addis Ababa University, 2011
10. T. H. Dipphone based Text- to- speech system for Tigrigna language. MSc, Addis Ababa University, 2011
11. Karjalainen M. Review of speech synthesis technology. Citeseer, 1999
12. Jayanthi K, Raghavan H, Murarka S, Murthy HA. Design and development of a text-to-speech synthesizer for Indian Languages
13. Borkar SP, Patil S. Text-to-speech System for Konkani (GOAN
14. Shreekanth T, Udayashankara V, Arun Kumar C. Unit selection based on hindi text-to-speech synthesis system using syllable as basic unit. IOSR J. 2014;4:49–57.
15. Nadew. Formant based speech synthesis for amharic vowels. M.sc, Computer Science Addis Ababa University 2008
16. Kishore SSP, Black AW, Kumar R, Sangal R. Unit selection voice for amharic using festvox. In: Fifth ISCA Workshop on Speech Synthesis, 2004
17. Tewdros. Text-to-Speech synthesizer for Wolayatta language. M.Sc., Information Science, Addis Ababa, 2009
18. Debela G. Afan Oromo news text summarizer. MSc., Information Science Addis Ababa. 2012
19. Roach P. English phonetics and phonology fourth edition: a practical course. Ernst Klett Sprachen; 2010.
20. Laver J. Principles of phonetics. Cambridge University Press; 1994.
21. Geshe D. Kamisee Oromoo phonology: one of the Oromo Dialects. MA, Addis Ababa University, 2010
22. Black A, Taylor P, Caley R, Clark R, Richmond K, King S et al. The festival speech synthesis system, version 1.4. 2. Unpublished document available via http://www.cstr.ed.ac.uk/projects/festival. html, 2001
23. Kayte S, Gawali B. A text-to-speech synthesis for Marathi language using festival and Festvox. Eur J Comput Sci Info Tech. 2015;3:30–41.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.