Unit Selection Based Text-to-Speech Synthesizer for Tigrinya Language

Agazi Kiflu
agazi_s@yahoo.com

Tibebe Beshah
School of Information Science, Addis Ababa University, Ethiopia
tibebe.beshah@gmail.com

Abstract

This paper brings together the development of the first unit selection based Text-to-Speech (TTS) system for Tigrinya using the Festival framework and practical applications of it. Construction of a unit database and implementation of the natural language processing modules are described and a Unit selection-based approach generates speech by selecting proper units from a speech corpus and connecting them together. In this approach, a set of features are defined to describe the speech units in the corpus and the expected units in the synthesized utterance. In this paper the major tasks have been performed, via development of concatenative Unit selection voice using phone as basic unit. We have used a speech corpus having a size of 4 hour, 38 minutes and 29 seconds, labelled at phoneme level.

We describe the implementation and evaluation of a G2P conversion model for a Tigrinya TTS system. Letter to sound conversion for Tigrinya usually has simple one to one mapping between orthography and phonemic transcription for most Tigrinya letters and an automatic clustering technique to cluster units based on their phonetic and prosodic context. Having constructed the phonetic, prosodic, and acoustic features extraction inventory for each phone to synthesize the input text the Festival speech synthesis was then adopted in order for the synthesizer to use cluster unit selection algorithm. In order to minimize acoustically defined target and join costs, a selection is made from cluster at the time of unit's synthesis.

The test results indicate that almost all of the words and sentences are recognizable. The system is evaluated using MOS, one of the most popular testing techniques in speech synthesis. The system is tested for naturalness and intelligibility of speech. On average, 97.1% of the sentences are correctly recognized by the listeners. The naturalness of the synthesized speech demonstrates the appropriateness of the proposed approach.

Keywords: Text-to-Speech Synthesis; Concatenative Speech Synthesis; Unit selection based speech synthesis; Syllable based concatenation; Consonants and Vowels

1. Introduction

Text-to-Speech (TTS) synthesis can convert arbitrary input text to intelligible and natural sounding speech so as to transmit information from a machine to a person [1]. It is a process through which input text is analyzed, processed, and understood and then the text is rendered as digital audio and then spoken [2]. The basic types of synthesis system are Formant, Concatenated, and Articulatory [3].

The process of TTS conversion allows the transformation of a string of phonetic and prosodic symbols into a synthetic speech signal. The quality of the result produced by a TTS synthesizer is a function of the quality of the string, as well as of the quality of the generation process. The most important qualities of a speech synthesis system are naturalness and intelligibility [2]. Naturalness describes how closely the output sounds like human speech, while intelligibility is the ease with which the output is understood. In this research, we used the concatenative text-to-speech system and the issues relevant to the development of a Tigrinya speech synthesizer using different choice of units: a word, phrase, clause, sentence or phonemes as a database. Since there is great advancement regarding TTS in other languages globally, attempt will be made to design and implement a method for TTS system in one of the local language of Ethiopia.

This paper is organized as follows: Section 2 focuses on the nature of the Tigrinya script. Section 3 explains literature review, while Section 4 outlines the methodology of the proposed solution with the

voice building process. Section 5 shows the results of perceptual testing and finally conclusion and recommendation are given in Sections 6 and 7.

2. The Tigrinya Language

The script of Tigrinya is phonetic in nature. It uses different choice of units: a word, phrase, has 39 consonants and 7 vowels [5, 9]. The orthographic representation of the language is organized into orders. Each of the 39 consonants has seven orders (derivatives). Six of them are CV combinations while the 7th is the consonant itself. The way Tigrinya orthographic characters are written is very similar to the way they are spoken. It means Tigrinya is a phonetic language. The mapping of the written form and the spoken form is one to one except the epenthetic vowel. Characters representing the same consonant followed by different vowels are similar in shape. For example, here are the characters representing: /he/, /hu/, /hi/, /ha/, /hie/, /h/ and /ho/: **ሀ ሁ ሂ ሃ ሄ ህ ሆ**.

The total number of orthographic Symbols of the language exceed 273. Like other languages, Tigrinya also has its own typical phonological and morphological features that characterize it. Among these, we found gemination of consonants and the use of the automatic epenthetic vowel to be very critical for naturalness in Tigrinya speech synthesis. Tigrinya language has special property in its spoken form (CV or CVC sequence of the acoustic form of the orthographic representation).

2.1 Phonology of Tigrinya Word

Phonology is the study of the distribution and patterning of speech sounds in a language and of the tacit rules governing pronunciation [4]. In phonology, phoneme is the fundamental unit that describes how speech conveys linguistic meaning. The phoneme represents a class of sounds that convey the same meaning. The meaning of a word is dependent on the phoneme that it contains [4].

Tigrinya has a fairly typical set of phonemes for an Ethiopian Semitic language. That is a set of ejective consonants and the usual seven-vowel system. Unlike many of the modern Ethiopian Semitic languages, Tigrinya has preserved the two pharyngeal consonants which were apparently part of the ancient Ge'ez language and which, along with [x'/ቐ]a velar or uvular ejective fricative, make it easy to distinguish spoken Tigrinya from related languages such as Amharic [9]. These are exception characteristics from Amharic beside their accent, manner and place of articulation

qa	k'	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
qwa	kw'	ቈ		ቍ	ቋ	ቌ	ቍ	
kxa	x	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ
kxwa	xw	ዀ		ዂ	ዃ	ዄ	ዅ	
qha	ʁ'	ቐ	ቑ	ቒ	ቓ	ቔ	ቕ	ቖ
qhwa	ʁw	ቘ		ቝ	ቛ	ቜ	ቝ	

Figure 1: Tigrinya Syllabic Structure

A syllable in Tigrinya is made up of only /**cv**/ and /**cvc**/ (a consonant + a vowel) or (a consonant + a vowel + a consonant). The vowel is a syllabic nucleus, while the first and the last consonants of the syllable are an onset and a coda respectively [9]. The Tigrinya native speaker, for example, can divide the words (sabara) and (biili) into three and two syllables respectively. For example, /qaatala/ (Figure 2) is one word which has three syllables in it. Some syllables of Tigrinya have a nucleus or peak and an onset, while other syllables have a coda in addition to an onset and a peak. Observe the following:

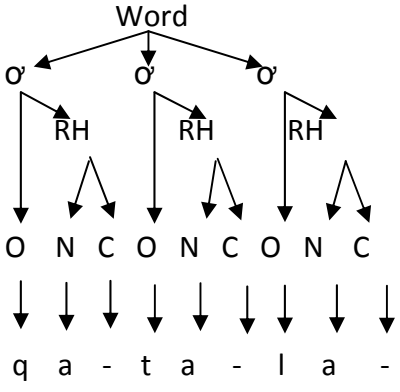

Figure 2

Moreover, each onset and coda position is occupied by a consonant, where as a nucleus position is occupied by a vowel.

2.2 Gemination

Longer duration of identical segments, adjacent consonants or vowels that are the same can form

germination. In Tigrinya sequence of vowels is not permissible. Whenever sequences of vowels occur, either one of the vowels must be deleted or epenthetic segments are inserted between the vowels. However, we do find geminated Tigrinya segments, with the exception of laryngeals and pharyngeals that may be geminated in only very limited environments as indicated in [9].

Consonant germination may bring meaning differences in words. If we compare /zawara/ 'he got roaming' and /zawwara/ 'he drove', /halifu/ 'he passed' and /hallifu/ 'he excelled'. There is a difference of meaning in each pair. In each pair, we observe a geminated or ungeminated medial consonant that brings a meaning difference in each of them.

2.3 Insertions

Insertion is one way for arriving at a well formed or acceptable assignment of syllable structure. The syllable structure of Tigrinya is either /cv/ or /cvc/. Insertion, unlike deletion, is the appearance of new elements in a formerly unoccupied position. The epenthetic (inserted) segments may appear word initially, word medially or word finally. There are several Tigrinya epenthetic segments, vowels and consonants, in different positions. Observe the following:

- asraha 'he made others to work'
- Awassaxom 'he made them to add'

The morpheme /a/ is added to the root consonants/srh/

- zii + asriih-a - zasriiha
- zii + awassaxa - zawassaxa

3. Literature Review

In [7] Indian Natural Language Processing Lab Centre for Development of Advanced Computing (CDAC) uses multiform speech unit to develop the speech synthesis. It primarily uses syllable and phonemes. The speech corpus contains most frequent words and initials. They have been segmented and labelled into different speech units as required for development of a Hindi speech synthesis system. This research doesn't indicate as to whether any kind of implementation has been made or not. Additionally, it is not known to have been tested or proven in any related manner.

Alam *et al.* in [8] proposed a TTS system that creates the voice data for festival, and additionally extends the use of festival to its embedded scheme scripting interface to incorporate Bangla language support. The researchers' TTS implementation used two different kinds of concatenative methods supported in Festival: unit selection and multi-syn unit selection [8].

The researchers on their future work indicated that a number of future plans need to be made to develop the complete TTS system for Bangla language including the following: document analysis, text analysis, phonetic analysis, developing large number pronunciation lexicon, automatic lexicon entries instead of adding manually, find out LTS or Grapheme-to-Phoneme (G2P) rule so that it can handle unknown words), prosody analysis, and waveform synthesis by diphone technique. In conclusion, the researchers observed that unit selection and multisyn unit selection has a drawback because of the requirement of large set of speech corpus.

Eker in [6] has found a research which exploits the Turkish language structure and tried to implement the system that takes a text as its input. It assumes that the text consists of words and it processes word by word [6]. When a word is obtained from the text, it is passed to a unit that can process word as text and produces the corresponding speech. This part separates the word into diphones; using diphone database, it gets a speech file corresponding to diphone and its pitch value.

Finally it concatenates the previously recorded speech segments using PSOLA algorithm and manages to produce sound. As a future work, the researcher recommended that the first thing should be done is to complete the diphone database and apply more experiments on words. The produced output is acceptable for small sentences, but it requires much time for long sentences. Therefore, in order to have a real-time reading system, the system should be faster. In conclusion, this means that the method used in this paper is an applicable one which with some effort on completing and preparing a

better diphone database will result in a system that will produce more understandable output for all Turkish words.

We have observed that few research attempts were made on local languages. One of the few attempts made was by Sebsibe H/Mariam *et al.* in [13]. Their focus was issues that need to be considered in developing a concatenative speech synthesizer. They have tried to describe the issues to be considered in developing a concatenative speech synthesizer for Amharic language. The complexity of the syllable structure of the language, the phonetic nature of the language, and the result of the perceptual test of the synthesizer has been discussed. The researchers tried to explore the nature of Amharic script representation of the phone set, Amharic syllable structure and syllabification rules, and showed the voice building process. Having noted that the quality of speech synthesiser for Amharic was not high, they recommended on the need in the future to work on improvement of the quality desired. They suggested that this can be done by:

1) Proper selection of unit. Since the language is phonetic, syllable as a basic unit may outperform the phone as a basic unit.
2) Optimal selection of corpus, which proportionally covers all basic units and variations, will give better quality.

Based on the reviewed made so far and knowledge of the researchers, none of the works so far have tried to design grapheme to phoneme converter or letter to sound speech synthesizer for Tigrinya. None of them show prototype for natural sound for Tigrinya, which synthesize by accepting normalized Tigrinya texts and generate prosodic features (i.e., intonation, stress) using syllable based approach. The main focus in this work is to find the proper quality speech corpus, which matches the quality of synthetic speech from synthesizers including linguistic tasks and develop naturally sounding text to speech for Tigrinya language.

4. Methodology

After an extensive literature review regarding concatenative speech synthesis method, unit selection concatenative synthesis is found to be the most popular method of performing speech synthesis recently and is found to differ from older types of synthesis by generally sounding more natural and spontaneous than formant synthesis or diphone based concatenative synthesis. Unit selection synthesis is proven to score higher than other methods in listener ratings of quality but it involves a tedious recording many hours of speech by a single speaker.

In this research we tried to explore the nature of Tigrinya script representation of the phone set, rules of letter to sound, Tigrinya syllable structure and syllabification rules that would show the voice building process. To do these researches we used the following.

- Transliteration scheme based on orthographic ordering of the script and acoustic similarity of the letters were defined using ASCII characters.
- In Festvox, the phone set of the language is described with the corresponding features like voicing, tongue position, tongue height, place of articulation, and manner.
- Experiments on Phonology of Tigrinya word, phone set and we try to cover all phonemes defined a transliteration scheme using ASCII characters.

5. Design and integration of Tigrinya unit selection into festival frame work

The speech inventory is divided into clusters, where each cluster holds units of the same phone class based on their phonetic and prosodic context. An outline of the steps to build a unit selection synthesizer are given below. A more detailed description of same is available in [10, 13].

- Design speech and text corpus
- Creating LTS rules and phone set
- Building utterance structures
- Generating speech unit clusters
- Building the unit synthesizer

What have been done in each step to build unit selection voice for Tigrinya is explained below. Tigrinya proverbs, articles, newspapers, magazine and bible sentence's are collected from different sources and are primary data. We selected a native

speaker of the language and tried to record in quite environment by a male speaker using PRRAT. We used wave surfer for manual labelling of the recorded voices. In this paper, we built a corpus of around 13171 words. The script of this speech corpus is selected from a large text corpus (around 84000 characters). The corpus is designed to cover the frequently used syllable and context as much as possible.

The input to the TTS system is the transliteration of a text in Tigrinya. The pronunciation generation module generates the sequence of basic units using a lexicon of units and letter-to-sound rules. The lexicon is a list of all speech units - monosyllables, bi-syllables and tri-syllables, present in the waveform repository. The letter-to-sound rules are framed in such a way that each word is split into its largest constituent syllable units As the pronunciation of most of the words in Tigrinya can be predicted from their orthography, these rules suffice to generate correct pronunciations. The unit selection algorithm generates a target specification for the speech units that have been identified and picks the best sequence of speech units that minimize both the target cost and the join cost. The waveforms of these speech units are then concatenated to produce synthetic speech.

5.1 System Architecture

As the system architecture shown in Figure 3, the synthesizer has text analysis and speech synthesis parts. The text analysis part uses grapheme to phoneme converter to match the word to its pronunciation whereas the synthesis part selects the best sequence of units for target specification produced at the end of text analysis, and finally generates the speech from of the speech parameters.

- Defining the phone-set of the language
- Tokenization and text normalization
- Incorporation of letter-to-sound rules
- Incorporation of syllabification rules
- Assignment of stress patterns to the syllables in the word
- Assignment of duration to phones
- Generation of f0 contour

Once a speech repository is in place, the repository is integrated with the Festival framework. An outline of the steps to build a unit selection synthesizer are given below. A more detailed description of same is available in [1, 2, 6, 10].

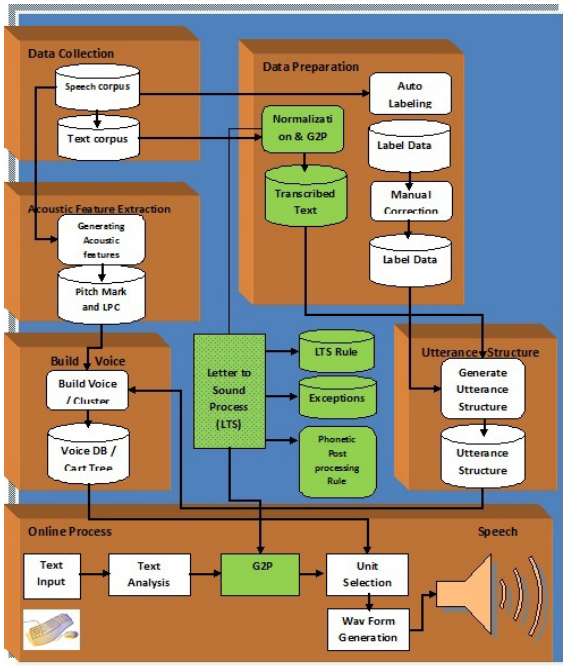

Figure 3: The system architecture of the Tigrinya speech synthesizer using cluster unit selection

5.2 Description of the Implementation Design

After we collected speech and text corpus, the next step was to check recorded utterances against the transcription text in order to design the prompt list in Festival format and correct the label manually. Appropriate modifications have been made to get them ready to be used in voice building process. By doing so the speech and text corpora has been built.

5.2.1 Labelling the Utterance

The process that generates the labelled utterance is labelling. Labelling is the process of giving a label for each speech signal in the utterance. Unit selection synthesizers are highly sensitive to the accuracy of labelling. Bad labels will adversely affect the quality of synthesis in a number of ways [2, 13].

The phone label itself can be incorrect, potentially causing the wrong word to be said, or said with an undesired accent. However, it is time taking and laborious, as part of our efforts to improve speech synthesis, we have labelled the speech database thoroughly using a tool called Wave Surfer.

5.2.2 Creating letter-to-sound rules and phone-set

A comprehensive set of letter-to-sound rules was created to syllabify the input text into the syllable-like units. These rules are framed in such a way that each word is split into its largest constituent syllable unit. The phone set, which is a list of basic sound units for Tigrinya that the synthesizer supports, was created by enumerating all the speech units identified in the syllabification process.

5.2.3 Incorporation of Tigrinya Phone set and Grapheme to Phoneme Converter

The phone-set definition is the first text analysis module in which every phoneme of the alphabet is classified according to phone features like consonant voicing and vowel height. The second text analysis module is the lexicon module.

The Tigrinya phone set is incorporated in Festival corresponding to their characterizing features. Each phone has eight features that describe how the vocal organs behave when the sound is uttered. These features are vowel/consonant identification, consonant voicing, place of articulation, consonant type, vowel length, vowel height, vowel front ness, and lip rounding.

The grapheme to phoneme converter is used to convert an orthographic text into its corresponding phonetic representation. After incorporation of Tigrinya phone set and the Tigrinya grapheme to phoneme converter into Festival, it provides the label files for each sentence in the prompt list. We have made manual label correction using the label automatically generated along with the corresponding wave file.

The grapheme to phoneme converter is used to convert an orthographic text into its corresponding phonetic representation. We implemented the grapheme to phoneme conversion architecture by making modification of syllabification algorithm proposed in [11]. A C# based syllabification program is implemented which is graphical based system and modified into C++ command line based G2P system is done as per the requirement of Festival tools. After incorporation of Tigrinya phone set and the Tigrinya grapheme to phoneme converter into Festival, it provides the label files for each sentence in the prompt list.

5.2.4 Building utterance structures for the database

The ‘utterance’ structure holds all the relevant phonetic and prosodic information related to a speech unit within this data structure. The phonetic information in an ‘utterance’ structure describes the position of the speech unit in the word it appears and the information of units adjacent to it. Prosodic information holds information about the duration and pitch of the unit. Festival provides relevant scripts for building ‘utterance’ structures for each speech unit.

5.2.5 Generating speech unit clusters

The process includes building coefficients for acoustic distances (MFCC, F0 and energy coefficients), creating distance tables for each class of units based on acoustic distances and generation of features for building CART trees.

5.2.6 Building the unit synthesizer

Using the letter-to-sound rules, phone set and clusters of each speech unit built in the previous steps, Festival generates the necessary files that need to be used along with the core Festival speech synthesizer to build a unit selection synthesizer for Tigrinya using appropriate scripts.

The rules of the language in relation to epithetic vowel insertion, gemination and syllabification effect on speech synthesis to determine the pronunciation of given Tigrinya words based on its spelling, in the process of grapheme to phoneme converter.

Epenthetic vowel insertion and germination rule for Tigrinya are adopted from [11] and modified with:

1. Accept input words and scan from left to right.
2. If consonant cluster occur at word initially position, insert epenthetic vowel between them.
   - Exception: If the first phoneme is consonant and the next consonant is glide/w/ pharyngeal/h/ plain/x/ (rule #1).
3. If three consonants are appeared in sequence word medially or word finally, position insert

epenthetic vowel before the third consonant (rule #2).

- Exception: if the middle consonant sonority is greater than the rest insert epenthetic vowel after next the first consonant.

4. If a cluster of consonant contains the germination and singleton in sequence, insert epenthetic vowel after the geminated consonant (rule #3).
5. If a cluster of consonant contains the singleton and geminate in sequence insert epithetic vowel after the singleton consonants (rule #4).
6. If a cluster of consonant contains two different germinations in sequence, insert epenthetic vowel between the two geminate consonants (rule #5).
7. If the sonority of the final consonant is greater than that of the proceeding consonants, the epenthetic vowel is inserted between the final consonant clusters (rule #6).
8. If a consonant cluster occurs at word final position, insert epenthetic vowel /i/.
9. Repeat steps 2 to 7 until the entire phoneme are parsed in the phonemes list.

6. Perceptual Evaluation and Experimental Results

Perceptual evaluation is essential to determine the quality of synthesized speech [13, 14]. The perceptual evaluation in this paper investigates the naturalness and intelligibility of Tigrinya TTS. In this research work mean opinion score (MOS) is used to test the output of the synthesized speech. MOS is an evaluation technique where evaluators indicate their assessments on a scale ranging from bad (1) to excellent (5). Then the average score of the opinion given will be taken as the performance of the system [6, 12, 14].

As we stated in the above section, the impact in perception is assessed by comparing the evaluation average result to be obtained from the score ranks given by native speakers at the end of their perceptual judgment for the synthetic speech produced by the synthesizer. Subsequently, the perceptual tests were carried out to evaluate the extent of naturalness and intelligibility of synthetic speech generated by the speech synthesizer.

6.1 Data Preparation and Prototype Testing

We conducted perceptua1 tests on 6 people who are native speakers of Tigrinya: 2 females and 4 males. All subjects are between 30 to 60 years old. Each subject listens to all of the 6 sentences with various lengths selected from the data set used in the voice construction and gives his/her ranking value for the naturalness and intelligibility of the speech. They evaluated based on the quality of the speech output by giving a measure of quality.

Based on the result found, we can conclude that proper selection of units done by the TTS has great role for perceived naturalness and intelligibility of synthetic speech sounds.

The results show that regarding the question as to whether the voice is good to listen to or not, 38.8% considered the voice is very good, 58.3% of them thought that the voice was good and 2.7 % considered the voice unnatural. From the result it is clear that more than 97.1% of the listeners found it to be ok and none of the listeners found it to be excellent, fair or very poor. In general, the output was acceptable by most of the listeners. When compared to a previous work done on unit selection in [8] for the Bangla language in Festival framework at sentence level, the average score is 90.1%, thus when compared to this thesis it has improved by 7.1% [8].

6.2 Summary

Even if the experiment is conducted on a small scale, the results obtained are promising. From this result, it appears to indicate that with an ever increasing size of speech database, the unit synthesizer would be able to produce natural speech with high flexibility and intelligibility. However, every feature of the Tigrigna language was not considered. This paper has achieved promising result by defining

- Transliteration scheme to work with Tigrigna scripts
- Incorporated phone set, Syllabification rules, and Letter to sound rules

- Stress assignment into Festvox.

7. Conclusion and Recommendation

In this work, a first attempt is made to develop a speech synthesizer for Tigrinya language using unit selection method. However, every feature of the Tigrinya language was not considered because it needs a lot of time and detailed and deep linguistic knowledge. Hence, only the characteristics and way of creation of Tigrinya phonemes are considered.

From the result and analysis, it can be concluded that this paper has achieved its main objectives. This project has produced a Tigrinya TTS system that has the ability to process the input of Tigrinya raw text to an output of Tigrinya speech sound. From the sentences level test also, it has proved that the user can understand almost every simple sentences spoken by the system.

The major contributions of this paper are:

- Identification of a new syllable based speech unit and suitable phone set for concatenative speech synthesis for Tigrigna,
- Development of LTS rule for Tigrigna.
- Development of natural sounding TTS systems for Tigrinya.
- Demonstration of the prototype.
- Investigation of techniques to develop speech synthesis systems.
- To assist raise further research question.

There are quite a lot of methods that can be used to improve this system. This improvement may range from its database method to its NLP processing method and the following points are recommended for future work either to extend the work or to increase the quality of the synthesized speech. As a future work we would like to suggest the following points:

- Syllabification of words: will greatly improve prosodic modeling with the segmental prosody of Tigrinya should be appropriately studied and modeled.
- The proper identification of Tigrinya stress point to help determine where should fall within a word.
- Automatic gemination and epenthesis handling algorithm.
- Deep studies on syllabification and final consonant cluster will improve the performance of the syllabifier.
- Stress assignment algorithm.
- Tigrinya morphological analyzer
- Duration modeling of consonants and vowels

References

[1] Thierry Dutoit, “A Short Introduction to Text-to-Speech Synthesis”, TTS research team, TCTS Lab.1997.

[2] A.W. Black, P. Taylor, and R. Caley, “The Festival speech synthesis system”, 1998.

[3] Juergen Schroeter, “Text to-Speech (TTS) Synthesis”, AT&T Laboratories, 2002.

[4] Nadew Tademe Mergia, “Formant based speech synthesis for Amharic vowels,” MSc Thesis, Faculty of Informatics, Addis Ababa University, Ethiopia, 2008.

[5] N. Sridhar Krishna, “Text-to-speech synthesis system for Indian languages within the Festival Framework:, M.S. Dissertation, Department of Computer Science and Engineering, Indian Institute of Technology, Madras, 2004.

[6] Barış Eker, “Turkish text to speech system”, MSc Thesis, The Institute of Engineering and Science of Bilkent University, 2002.

[7] Natural Language Processing Lab Centre for Development of Advanced Computing, “Building speech corpora for unit selection based concatenative text to speech system for Indian Languages”, India.

[8] Firoj Alam, Promila Kanti Nath, and Mumit Khan, “Text To Speech for Bangla Language using Festival”, BRAC University, Bangladesh.

[9] Tesfaye Tewolde Yohannes, “A modern Grammar of Tigrinya,” Rome, Italy, 2002.

[10] Alan W Black and Kevin A Lenzo, “Building Synthetic Voices, For FestVox 2.0 Edition”, Language Technologies Institute, Carnegie Mellon University and Cepstral, LLC, 2003.

[11] Nirayo Hailu Gebregziabher “Modeling Improved Amharic Sylliblification Algorithm”,

MSc Thesis, Faculty of Informatics, Addis Ababa University, Ethiopia, 2011.

[12] Sebsibe H/Mariam, S P Kishore, Alan W Black, Rohit Kumar, and Rajeev Sangal, "Unit Selection Voice for Amharic using Festivox", 5th ISCA Speech Synthesis Workshop, Pittsburgh, pp. 103-107, 2005.

[13] Yonas Demeke, "Duration modeling of phonemes for Amharic text to speech system", M Sc Thesis, Faculty of Informatics, Addis Ababa University, Ethiopia, 2011.

[14] Hyunsong Chung, "Duration Models and the Perceptual Evaluation of Spoken Korean", Proceedings of ISCA Archive, France, 2002.

[15] AlanW Black and Kevin A. Lenzo, "Building Synthetic Voices", For FestVox 2.1 Edition. 2007.