Global Scientific JOURNALS

GSJ: Volume 8, Issue 6, June 2020, Online: ISSN 2320-9186
www.globalscientificjournal.com

Developing Concatenative Based Text To Speech Synthesizer for Tigrigna Language

Mezgebe Araya Keletay[1, 2*], Hussien Seid Worku[2]

1. **Department of Computer Science, School of Computing and Informatics, Mizan-Tepi University, Tepi,Ethiopia**
2. **Department of Computer Science, College of Engineering and Technology, Arba-Minch Institute of Technology, ,Arba-Minch,Ethiopia**

Email address:
fanyene23@gmail.com (Mezgebe Araya Keletay), lehussien@gmail.com (**Hussien Seid Worku**)

***Corresponding Author**

Abstract

A Text-To-Speech (TTS) synthesizer is a computer-based system able to read any text and convert it into speech that resembles as closely as possible a native speaker of the language. This thesis describes the first Text-to-Speech (TTS) system for the Tigrigna language, using speech synthesis architecture in MATLAB. The TTS system is working based on concatenative synthesis and applying LPC technique. The performance of the system is measured and the quality of synthesized speech is assessed in terms of intelligibility and naturalness. The result of the synthesizer is evaluated in two ways, in word level and sentences level. The test results indicate in the word level is evaluated by NeoSpeech tool online and most of the words are recognizable. The overall performance of the system in the word level which is evaluated by NeoSpeech tool is found to be 78%. When it comes to the intelligibility and naturalness of the synthesized speech in the sentence level, it is measured in MOS scale and the overall intelligibility and naturalness of the system is found to be 3.28 and 3.27 respectively. The values of performance, intelligibility and naturalness are encouraging and show that diphone speech units are good candidates to develop fully functional speech synthesizer. But there are areas that can be improved. Inclusion of text analyzer to pronounce zonal dialects of the language and prosody generator are some of the things that need further investigation.

Keywords: Concatenative approach; speech synthesis; Tigrigna syllables; Text-to-Speech

1 Introduction

Language is a fundamental part of everyday life human being. Whether we are using speech, sign language, emotion or a coding system that conveys meaning through touch, we use language to express our thoughts, intentions, reactions, and experiences [1].Text-to-speech (TTS) synthesizer transforms linguistic information stored as data or text into speech. It is most widely used in the audio reading devices for the visually impaired people now days. TTS is one of the major applications of NLP. The NLP module of general TTS synthesizer consists of the Pre-processor, text

analyzer, contextual analyzer [2], syntactic prosodic parser, letter to sound module and prosody generator. Synthesized speech can be created by concatenating part of recorded speech which is stored in a database. Speech is often based on concatenation of natural speech that is the units, which are taken from natural speech put together to form a word or sentence [3].

Text-To-Speech (TTS) synthesis system has a wide range of applications in everyday life. And a text to speech synthesizer is used for vocalization processed content [4]. In last decade, a great deal of TTS-Synthesis system has done much work in various languages as well as different synthesis techniques such as Unit-selection, Formant, Hidden Markov Model and Articulatory synthesis was done by researchers [4].In order to make the computer systems more interactive and helpful to the users, especially physically and visually impaired and illiterate masses, the TTS synthesis systems are in great demand for the Ethiopian languages [5].

Research in the area of speech synthesis has been worked by the growing importance of many new applications. These include information retrieval services over telephone such as banking services, public announcements at places like train stations and reading out manuscripts for gathering [7]. Speech synthesis has also found applications in tools for reading emails, faxes and web pages over telephone and voice output in automatic translation systems. Special equipment for the physically challenged, such as word processors with reading-out capability and book-reading aids for visually challenged and speaking aids for the vocally challenged also use speech synthesis [8].

© GSJ

The growing popularity of speech-enabled computer interfaces demands high quality speech output, particularly for telephone applications. The perceived quality of standard general purpose text-to speech (TTS) systems is not good enough, which forces application developers to use pre-recorded prompts, drastically reducing the text generation flexibility. Recent improvements in limited-domain synthesis have been in the context of concatenative synthesis, with a focus on methods for combining whole phrases and words with sub word units for infrequent or new words. Little or no attention has been paid to natural prosody generation, with the assumption that it is accounted for in the phrase-size units. However, as complexity of the domain increases, there is more room for prosodic variability that must be accounted for to achieve natural speech [9].

1.1 The Tigrigna Language

Tigrigna, often written as Tigrinya (ትግርኛ) is a language spoken in the east African countries such as Eritrea and Ethiopia. It is one of the two official languages of the country Eritrea. It is also a working language of the Tigray region of Ethiopia. According to the 2015 Census conducted by the Agency of Ethiopia (CSA), the Tigray Region has a population of 6.3 million and from the total population around 4.3 million are native Tigrigna speakers, and according to Ethnologies there are 2.4 million Tigrigna speakers in Eritrea [10].

The script of Tigrigna is phonetic in nature. It has 35 consonants and 7 vowels [6]. The orthographic representation of the language is organized into orders. Each of the 35 consonants has seven orders (derivatives). Out of the 35

consonants four of them are diphthongs. Six of them are CV combinations while the 7th is the consonant itself. The way Tigrigna orthographic characters are written is very similar to the way they are spoken. It means Tigrigna 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 [6].

Table 1: Tigrigna Syllables structure of character "ሀ" and "ለ"

	1st	2nd	3rd	4th	5th	6th	7th
ሀ	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
	He	Hu	Hi	ha	Hie	h	ho
ለ	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
	Le	Lu	Li	La	Lie	L	Lo

1.2 Characteristics of Tigrigna Language

As same with other Semitic languages, Tigrigna has its own characterizing phonetic, phonological, and morphological properties. Tigrigna language has its own characterizing phonetic, phonological and morphological properties. It has a set of speech sounds that is not found in other languages. For example the following sounds are not found in English and Amharic: [ʔ](ዐ), [ħ] (ሐ), [k'] (ኸ), [ʔ] (አ) and [x'] (ቐ) [10].Tigrigna also has its own inventory of speech sounds. Fidel's (alphabets) have the same pronunciation but different symbols, these different Fidel's can be used interchangeably without meaning change. The Fidel's are "ጸ" and "ፀ", "ሰ" and "ሠ" and "ሀ", and "ኀ". For example, the word "Hair" can be written as, "ጸጉሪ", "ፀጉሪ", the word "weed" can be written as, "ፀሃየ", "ፀሓየ", "ጸሃየ", and "ጸሓየ" ,the word "hunter" can be written as, "ሃደነ", "ሓደነ", and the word "troop" can be written as, "ሰራዊት", "ሠራዊት" etc, all mean the same, although they are written differently and produce different orthographic form.

1.3 Consonant Phonemes

There are thirty-five consonant phonemes in Tigrigna .The consonants are generally classified as Stops, fricatives, nasals, liquids, and semi-vowels. Unlike many of the modern Ethiopian Semitic languages, Tigrigna has preserved the two pharyngeal consonants which is apparently part of the ancient Ge'ez language and which, along with [x'], which is "ቐ", a velar or uvular ejective stop make it easy to distinguish spoken Tigrinya from related languages such as Amharic. The fricative sounds [x], which is "ኽ", [x^w], which is "ኾ", [x'] which is "ቐ", and [$x^{w'}$] which is "ቖ" occur as allophones [6].

Table 2: Tigrigna Syllabic Structure

Kxa	**x**	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ
Kxwa	**xw**	ኾ		ኹ	ኻ	ኼ	ኾ	
Qa	**k'**	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
Qha	**b'**	ቐ	ቑ	ቒ	ቓ	ቔ	ቕ	ቖ
Qhwa	**bw'**	ቘ		ቖ	ቚ	ቜ	ቖ	
Qwa	**kw'**	ቈ		ቊ	ቋ	ቌ	ቍ	

1.4 Vowel Phonemes

Vowels are always voiced sounds and they are produced with the vocal cords in vibration [1]. Most languages have five vowels/a, e, i, o, u/, but in case of Tigrigna, there are seven vowels. These are ኣ, ኡ, ኢ, ኣ, ኤ, እ, and ኦ. All are voiced and oral sounds. These vowels can be found in each letters, that is, each letter in Tigrigna is not a single sound rather they are a combination of two sounds, one from vowel and one from consonant. Depending on the position of the lip the Tigrigna vowels (ኣ፡ኡ፡ኢ፡ኣ፡ኤ፡እ፡ and ኦ) are broadly categorized into rounded (ኡ and ኦ) and unrounded (ኣ፡ኢ፡ኣ፡ኤ and እ) [1].

1.5 Gemination

Gemination/ጥበቀት/ (consonant lengthening) is not normally indicated in the Ge"ez script. Longer duration of identical segments, adjacent consonants or vowels that are the same can form in Tigrigna sequence of vowels is not permissible. Consonant gemination may bring meaning differences in words. If we compare "ዘዋሪ" /zawara/ "he got roaming" and "ዘዋሪ" /zawwara/ "he drove", and the word "ሓሊፉ" /halifu/ „he passed" and "ሓሊፉ" /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 Literature Review

Speech synthesis is the processes of converting a written text into speech and this technology have the ability to convert arbitrary text into audible speech, with the goal of being able to provide textual information to people via voice messages [11] . The speech synthesizer depends on the TTS synthesizer architecture inculcated to produce intelligible and natural sounds from the synthesizer.

Figure 1: Text to Speech System

2.1 The Natural Language Processing (NLP) Component

Natural Language Processing or text-to-phoneme (T2P) is targeted to produce phonetic transcription of the text, together with the desired prosodic features [9]. It concern how computational methods can aid the understanding of human language and focused on developing systems that allow computers to communicate with people using every day in their life. The components are text analysis, automatic phonetization and prosody generation [1].

There are a number of factors which is affected natural language processing and the final output of digital signal processing. Some of the factors which affected in this research works like, environmental affects during record time, quality of microphone, sampling frequency, echo and noise.

2.2 The Digital Signal Processing (DSP) Component

The digital signal processing unit transforms the symbolic information that receives from NLP into audible and intelligible speech. Automatically, the operations involved in the DSP component are the computer analogue of dynamically controlling the articulatory muscles and the vibratory frequency of the vocal folds so that the output signal matches the input requirements [1].

© GSJ

2.3 Speech Synthesis Techniques

Synthesized speech can be produced by employing several different techniques to find natural human like sounds. The main techniques of speech synthesis synthesizer are discussed below:

2.3.1 Articulatory Synthesis

Articulatory synthesis tries to model the human speech production system (especially vocal tract system, various articulators like, Lip, tongue, jaw etc…) and articulatory processes directly. However, it is also the most difficult method to implement due to lack of knowledge of the complex human articulation organs [12].

2.3.2 Formant Synthesis

Formant synthesis is based on the rules which describe the resonant frequencies of the vocal tract. The formant method uses the source-filter model of speech production, where speech is modeled by parameters of the filter model. Rule-based formant synthesis can produce quality speech which sounds unnatural, since it is difficult to estimate the vocal tract model and source parameters [5].

2.3.3 Unit Selection Synthesis

Unit selection based Concatenative speech synthesis, joint cost also known as Concatenative cost, which measures how well two units can be joined together [13].

2.3.4 Concatenative Synthesis

Systems can synthesize high quality and more natural sound speech but in order to synthesize speech with various voice characteristics such as speaker individualities, speaking styles, emotions, etc., a large amount of speech corpus and memory is required as stored basic speech units (like syllables, diphones etc.) are concatenated to form word sequence using pronunciation dictionary [13].

Concatenative synthesis is concatenating the pre-recorded segments to generate the natural speech. Concatenative speech is produce intelligible & natural synthetic speech, usually close to a real voice of person [13]. However, concatenative synthesizers are limited to only one speaker and one voice. The difference between natural variation in speech signals and the nature of the automated techniques are segmenting the waveforms form the audible output [14].

© GSJ

3 Methodology

Research methodology is the process of used to collect information and data for the purpose of making decisions regarding of the research title. Research methodology may include publication researches, interviews, surveys and other research techniques are used.

3.1 Research Strategy

The research thought with respect to this thesis work was an applied one, but not new. Somewhat, numerous researches are existing regarding the role of TTS in different local and international languages to synthesis the natural languages automatically for the purpose of minimizing the challenges in day to day activities specially visual impaired peoples, not only for Impaired peoples in specific, but also for non-blinded peoples are also usable.

3.2 Research Approach

There are different approaches to develop a text to speech synthesizer, such of the approaches are discussed in chapter two, but this research was used a concatenative based approach to synthesis the Tigrigna TTS model. In concatenative approach which records the Tigrigna diphones (half phone) which is known as “Fidels”. The prerecorded sounds of Tigrigna were concatenated to get a words, phrases, and sentences of Tigrigna using a concatenative approach. The systems in concatenative approach can synthesize high quality and more natural sound speech was listened by the native speakers of Tigrigna language.

3.3 Data Collection Method and Tools

The direct observation and review of articles are applied in this research paper to identify the whole strings which is represented the language (the "Fidels") and tools used to develop and test the TTS synthesizer respectively. Tools which are used in this research paper was PRAAT, which is used to record and analyze the strings ("fidels") of Tigrigna language, MATLAB was used to implement the Tigrigna TTS synthesizer, and Neospeech was used to test the performance of the TTS synthesizer.

3.4 Data analysis

Data analysis is a content analysis which is used to analyze the data which was gathered from interviews and direct observations. Therefore, in this research work the gathered information's are analyzed using a tool of praat. The gathered data or the strings ("Fidels") of Tigrigna language are collected from spiritual notes of Geez scripts which is known as "Abugida" and the collected strings are recorded and analyzed using PRAAT. Natural sounds are collected from different articles, journals, and newspapers of Tigrigna language and analyzed to phones, words, phrases, and sentences to check the performance evaluation of the TTS synthesizer.

© GSJ

3.5 Research Method

The research methodology provides an orientation that influences the research results, procedures, evaluating validations of the research work. Tigrigna corpus was prepared to implement a TTS synthesizer using the tool of PRAAT by recorded the Tigrigna diphones in wav file. Then after the recorded wav file phones are changed to txt files using the tool of MATLAB. Subsequently, the txt file is read automatically in the MATLAB and linear productive coding (LPC) was applied to estimate the error signals in order to get the natural sound. Then, the TTS synthesizer was checked its performance in two techniques, the first one is by using the tool of NeoSpeech in order to test the sample words of their naturalness and intelligibility of the synthesizer. Secondly, the mean opinion score (MOS) was used to test the sample sentences by invited 20 native speakers of the language.

Finally, the overall result using diphones to synthesize Tigrigna language with 78% accuracy and the overall intelligibility and naturalness of the system from twenty listeners for the ten Tigrigna sentences is found to be 3.27 and 3.28 respectively.

3.6 Sample Selection

The method of sampling was used to develop the sample of the research under discussion. According to this method, which belongs to the sampling size, are selected on the basis of implemented the TTS synthesizer, evaluated the performance of the TTS synthesizer and testing the TTS Synthesizer. In this research work 35x 35

diphones are recorded to develop the TTS model for Tigrigna language. Additionally, to test the TTS model 100 Tigrigna words and 10 different sentences were used and to check the performance of the synthesizer twenty (20) native speakers are participated, out of them 12 persons are men and the remaining 8 persons are women.

4 Design an Automatic Model Text to Speech Synthesizer for Tigrigna

The demonstration of text to speech synthesizer model is how it could be designed, implemented and integrated the input texts matching with its database. Algorithms enable to modify the pitch and duration of the speech to achieve synthesized speech by concatenating diphone segments.

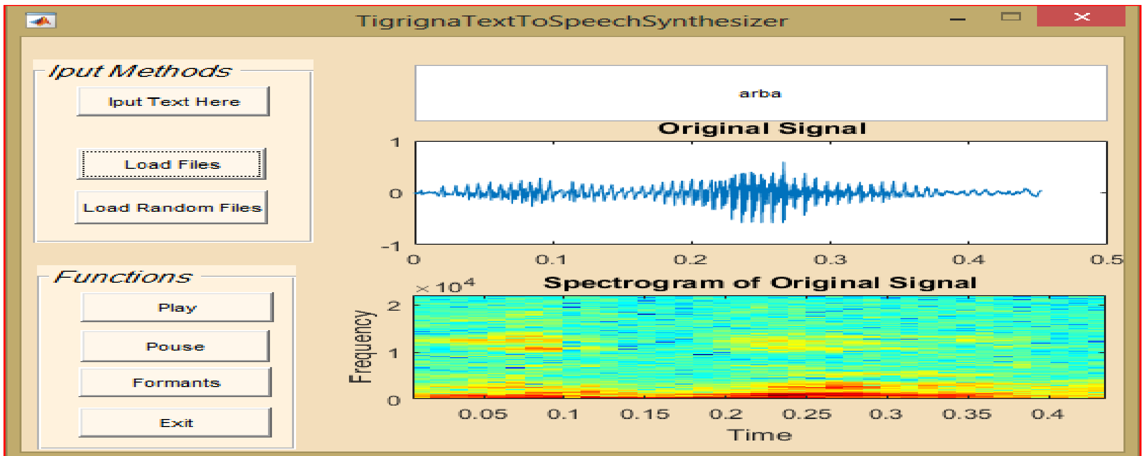

Figure 2: GUI Text-to-Speech Synthesizer for Tigrigna

4.1 Linear productive coding

Linear productive coding is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form using the information of a linear predictive model [15].There are various advantages for the use of LPC and they are.

a. LPC proves better approximation coefficient spectrum
b. LPC gives shorter and efficient calculation time for signal parameters and
c. LPC has been able to get important characteristics of the input signals.

$$S[n] = \sum_{k=1}^{n} (a_k \, S([n-k]) \quad \ldots\ldots 1$$

Where P is the number of past samples of s[n] which we wish to examine.

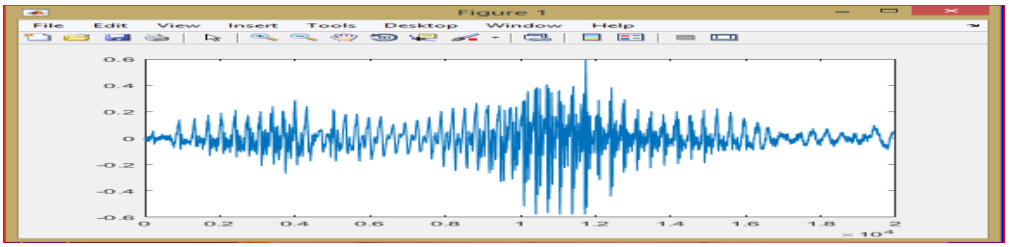

Figure 3: The original signal of the word "arba"

The algorithm which is used to read files from the database in concatenative approach is as follows:

1) for check the text file from one to N
2) Load text file from database
3) concatnate one to N
4) read the text
5) End for

ALGORITHM 1(steps to read a file):

STEP1: Create a database of various wave files

STEP2: Create a text file (.txt)

STEP3: Open the .txt file in matlab.

STEP4: Read the file opened.

STEP5: For every character read, play the corresponding wave (.wav) file.

© GSJ

4.2 Proposed Architecture of Text to Speech Synthesizer for Tigrigna

Basically there are three main modules that are used to build TTS synthesizer for Tigrigna: the Natural Language processing module, the Digital Signal Processing Modules and the Database modules.

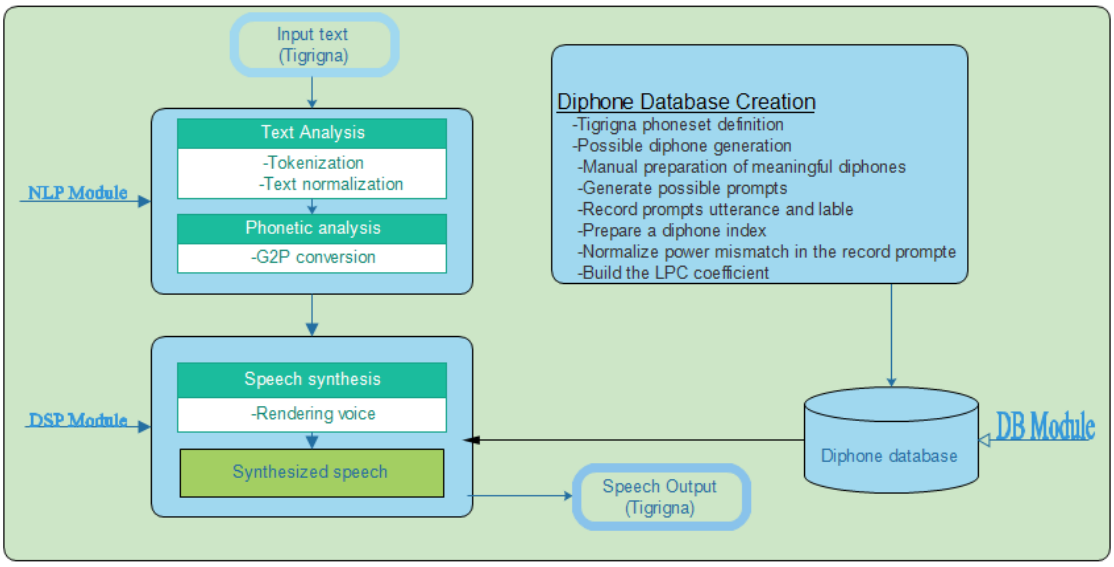

Figure 4: TTS Architecture for Tigrigna

4.3 Experimental results and Discussions

The first experiment is on the performance of the system that is assessed on word level. The test consists of 100 Tigrigna words selected through the help of a native speakers of the language. The selected words are evaluated their naturalness and intelligibility using a software tool called NeoSpeech. Therefore, the researcher gives the selected words for the tool and listen their naturalness and intelligibilities of the sound which is played by the tool online.

The overall performance of the system is measured in terms of total number of correctly pronounced words over the total number of words played. Finally by calculating the number of words which are correctly pronounced the overall performance of the system is found to be 78%.

The second experiment evaluated intelligibility and naturalness of the synthesizer. In this research Mean Opinion Score (MOS) technique is used to evaluate the synthesized text because it is the most widely used and simplest method to evaluate speech quality [6].

The overall intelligibility of the system from twenty listeners for the ten Tigrigna sentences is found to be 3.27. Which means the synthesizer is ‘good’ as per the scale of the MOS test. The overall naturalness of the synthesizer found to be 3.28 which also approach to ‘good’ MOS scale.

© GSJ

5 Conclusion and Recommendation

5.1 Conclusion

Text-To-Speech (TTS) synthesizer is a computer-based system that should be able to read any text aloud, when it was directly introduced in the computer by an operator.

Text Analysis which is capable of converting raw text to pronounceable words, Phonetic Analysis which converts text in orthographic form to phonemes, certain properties of the speech signal are processed, Diphone database Creation which provides diphone speech units to be concatenated and uttered and Diphone Concatenation where the speech is generated.

Based on the evaluation, the system register on the average 78% performance; 3.28 MOS score in intelligibility and 3.27 MOS score for naturalness. The result looks encouraging and further improvement of intelligibility and naturalness depend on proper works in different context. In this research we prepared diphone inventory in consultation with the domain experts. But as proved in different literatures having well studied diphone units produce better quality sound.

5.2 Recommendation

Based on the findings of the study, we recommend the following to improve the quality of the system and to enhance the quality of the synthesized speech.

In this study we did not consider prosody, word stresses, intonations and zonal dialects of the language, which are challenging in designing the speech synthesis.

Speech emotion development for different type emotions like normal, happy, anger, and sad, fear and grief are some of the emotion type which make the speech output as well as waveform generation varied. Therefore, there is much work that could be carried out in this area alone. However, future work in other emotions may not produce the same results found in this thesis. This would be due to a number of reasons: more complex emotions are less understood and as a consequence of speech correlates for complex emotions are much harder to identify.

Acknowledgements

The corresponding author would like to thank the Department of Computer Science in Arba-Minch University to support and advice worth considering starting from the beginning to the completion of the paper, and the native speakers of the Language they supports me to give their interests for recording. Next I would like to thank my advisor, Dr. Hussien Seid for his tireless support, patience, guidance, and encouragement.

Reference

[1] M. S. SIYOUM, "SYLLABLE-BASED TEXT-TO- SPEECH SYNTHESIS (TTS) FOR AMHARIC," ADDIS ABABA UNIVERSITY, June, 2012.

[2] R. K. Kaveri Kamble, "Translation of Text to Speech Conversion for Hindi Language," 2012.

[3] S. A. S. S. P. P. Mrs. Mangal Joshi, "Text to Speech Synthesis for Hindi Language using Festival Framework," *International Research Journal of Engineering and Technology (IRJET),* vol. 06, no. 04, p. 630, Apr 2019.

[4] Dr. Samuel Manoharan, "A SMART IMAGE PROCESSING ALGORITHM FOR TEXT RECOGNITION, INFORMATION EXTRACTION AND VOCALIZATION FOR THE VISUALLY CHALLENGED," *Journal of Innovative Image Processing (JIIP),* vol. 01, pp. 31-38, (2019.

[5] R. J. R. G. D. Ramteke, "Text-To-Speech Synthesis of Marathi Numerals," vol. 3, no. 7, July 2015.

[6] A. Kiflu, "Unit Selection Based Text-to-Speech Synthesizer for Tigrinya Language," vol. Volume 1, December 2012.

[7] A. T. Ei Phyu Phyu Soe, "Text-to-Speech Synthesis for Myanmar Language," *International Journal of Scientific & Engineering Research,* vol. 4, no. 6, p. 1509, June 2013 .

[8] J. M. Varghese, "Design of Gujarati Text-to-Speech System," vol. 02 , no. 05, May 2015.

[9] B. Sudhakar, "Development of Concatenative Syllable based Text to Speech Synthesis System for Tamil," vol. 91, April 2014.

[10] Y. FISSEHA, "DEVELOPMENT OF STEMMING ALGORITHM FOR TIGRIGNA TEXT," JUNE 2011.

[11] N. P. P. S. S. a. S. A. Ayushi Trivedi, "Speech to text and text to speech recognition systems-Areview," *IOSR Journal of Computer Engineering (IOSR-JCE),* vol. 20, no. 2, p. 40, May-April 2018.

[12] S. D. D. E.Kodhai, "Textaloud Assistant App Development for Multilanguage," *International Journal of Innovative Technology and Exploring Engineering (IJITEE) ,* vol. 8, no. 7s, May 2019 .

[13] M. R. B. ,. C. N. M. Suhas R. Mache, "Review on Text-To-Speech Synthesizer," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 4, no. 8, p. 56, August 2015.

[14] P. G. K. D. Pawan S. Nadig, "Survey on text-to-speech Kannada using Neural Networks," *International Journal of Advance Research, Ideas and Innovations in Technology,* vol. 5, no. 6, p. 128, 2019.

[15] G. D. R. R. J. R. Sunil S. Nimbhore, "Implementation of English-Text to Marathi-Speech (ETMS) Synthesizer," *IOSR Journal of Computer Engineering (IOSR-JCE),* vol. 17, no. 1, pp. 34-43, Feb. 2015.

[16] Y. B. Ilyes Rebai, "Text-to-speech synthesis system with Arabic diacritic recognition system," *Multimedia InfoRmation System and Advanced Computing Laboratory,* 17 April 2015.

[17] A. T. ZEGEYE, "A GENERALIZED APPROACH TO AMHARIC TEXT-TO-SPEECH (TTS) SYNTHESIS SYSTEM," Addis Ababa University, July, 2010 .

© GSJ