Innovative Systems Design and Engineering
ISSN 2222-1727 (Paper) ISSN 2222-2871 (Online)
Vol.4, No.9, 2013-Special Issue - 2nd International Conference on Engineering and Technology Research
www.iiste.org

Development of Text to Speech System for Yoruba Language

Akin Afolabi [1*] Elijah Omidiora[2] Tayo Arulogun [3]
Department of Computer Science and Engineering, Ladoke Akintola University of Technology Ogbomosho Nigeria
* E-mail of the corresponding author: akinloluade@gmail.com

Abstract
Text-to-speech (TTS) applications have been applied to different languages in diverse areas of human endeavour all over the world, but for Yoruba language which is being spoken by over 30 million people out of 150 million Nigerian populace and in other countries like Benin, Togo, United Kingdom and part of South American, much has not been achieved therefore, there is need to develop TTS system for Yoruba language. This paper gives an account of Yoruba TTS system development using concatenation method. The paper describes the design, evaluation and the analysis of the result shows that 70% Respondents accepted its usability.
Keywords: TTS, evaluation, concatenation, usability and design.

1. Introduction

To many people, the term speech synthesis evokes memories of mechanical, monotonous or repetitive voices but what really is a Text-to-Speech system? It is simply defined as a written text transformed into speech, it may be by reading or dictating through machines, input is text and the desired output is an acoustic speech signal, therefore comes the name text-to-speech synthesis.
There are two major types of TTS, these are Parameterised and Concatenative, Parameterised TTS can be further categorised into Formant base and Articulatory TTS. Formant based uses rules based on signal from the spoken input while Articulatory TTS make use of model of the vocal tract based on electro-acoustics theories. Concatenative synthesis makes use of mathematical model based on phonemes or syllables and produces a speech fragments as its output. The resulting speech was slightly artificial, sound like the original speaker who his or her voice was taking as sample. This type requires a powerful algorithm and larger memory capacity because each unit of speech needs memory space.
In the last few years, this technology has been widely available for several languages for different platform ranging from personal computer to stand alone systems. Though for Yoruba language in Nigeria, TTS has not been fully developed by the researches except a Standard Yoruba TTS System (Odejobi, 2006) and web-based aided tutor for Yoruba Language (Odetunji, 2003). Though the research in this field is still going on with African Languages Technology Initiative (Adegbola, 2011) but much has not been achieved. For this paper Concatenative Method was preferred because it produces a very close to humanlike voice and because of distinct features of the language with this method can accommodate e.g. tone, syllabic stinging.

2. Review of related works

The modern TTS system converts text into 'synthetic speech sound in a two-stage process (Klatt, 1976). The first stage i.e. High Level Synthesis (HLS) reads the input text and generates a representation of how the text will be pronounced. The HLS stage is implemented using two modules, the first module, i.e. Text-analysis module, analyses the input text to identify its basic elements and the context in which they are used.
The results of the text-analysis module is fed into the second module i.e. prosody module, which generate a linguistic description of how the text will be pronounced. It also integrates timing and rhyme information into the generated representation. All the processing involved in this stage are together called High level Synthesis (HLS) and the technology for implementing them were draw from the domain of Natural Language Processing (NLP) and computational Linguistic (Sproat, Black, Chen, Kumar, Ostendsorf and Richards.1996). A TTS system is composed of two parts a front-end and a back-end. (Van, Richard, Joseph and Julia, 1997). The front-end has two major tasks. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words. This process is often called text normalization, pre-processing, or tokenization. The front-end then assigns phonetic transcriptions to each word and divides, and marks the text into prosodic units, like phrases, clauses and sentences (Van et al., 1997).
Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer and can be implemented in software or hardware (Allen, 1987). High level Synthesis method was used to develop a TTS for Yoruba Language by Odejobi (Odejobi et al.,2006).
The major concern of any TTS system is to ascertain the intelligibility and naturalness of the synthesizer and this is achievable based on the type of method use in the designing of the speech synthesizer system (Afolabi, 2012). The focus of this paper is to concatenate some Yoruba syllables to produce a speech.

2.1 Yoruba Phonology

The Yoruba alphabet consists of 25 letters which were derived from Latin characters. (Adewole, 1988). 18 out of the 25 alphabet constituted Yoruba consonant and the remaining 8 were vowel this was illustrated in Table 1 and 2.

Table 1. The upper case and lower case representation of Yoruba alphabet

Aa	Bb	Dd	Ee	Ẹẹ
Ff	Gg	GBgb	Hh	Ii
Jj	Kk	Ll	Mm	Nn
Oo	Ọọ	Pp	Rr	Ss
Ṣṣ	Tt	Uu	Ww	Yy

Table 2. Orthography Representation of Yoruba consonant

Bb	Dd	Ff	Gg	GBgb
Hh	Jj	Kk	Ll	Mm
Nn	Pp	Rr	Ss	Ṣṣ
Tt	Ww	Yy		

2.2 Syllable Structure

The three basic syllable types in Yoruba are Vowel (V), Consonant-Vowel (CV) and Nasal (N). The first type of syllable involves only a single vowel and this is often the shape of lexical items such as pronouns. The second syllable type in Yoruba is a consonant and a vowel this is the basic shape of simple verbs in the language. The third and final syllable type in Yoruba is the syllabic nasal. Due to the shape of the syllable types in Yoruba, there are no consonant-final words and therefore, there are no closed syllables in the language. All the three syllable types have either combinations of first and third type or combinations of first and second and it can be vowel only (Nucleus) The three syllable types are illustrated in the Table 3.

Table 3. Yoruba syllable types

Syllable	Example	Meaning
Vowel (V)	a ó	‘we’ 'he'
Consonant Vowel (CV)	r án t à	‘to sew’ ‘to sell’
Nasal Vowel [N]	ò ro n bó dù n dú	‘lemon ‘fried yam’

2.3 *Supra Segmental Elements*

Yoruba is a tonal language, it has three surface tones of different pitch levels. The syllable is the tone bearing unit in the language but orthographically, tones are marked on vowels and syllabic nasals (Odejobi, 2006.)

3. Architectural Design

Models are of many types which include: iconic, graphical, simulation, textual mathematical and axiomatic models. Graphical model approach was used in this work. For this research, Thierry Dutoit model was adapted.

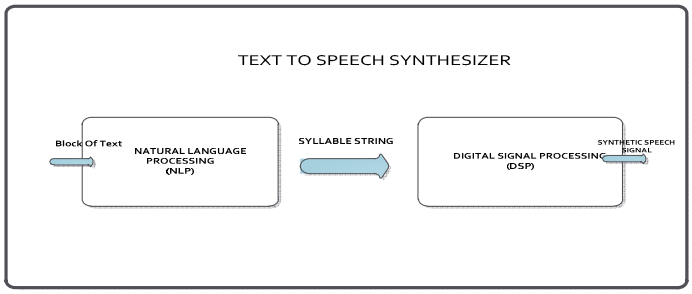

Figure 1: General Structure of TTS (adapted from Thierry Dutoit 2000)

The components of a TTS consist of two major components they are Natural language Processing (NLP) and Digital Signal Processing (DSP). NLP component performs the task of decomposing a sentence to its sequence of the parts of speech such as noun, verb, adverb etc. It consists of Text Analyzer which tokenizes a block of text and Syllable Analyzer which deduce the syllable to be use for specific word. The block of text will be fed in and it will be processed by parsing through the text after this, block of text will be tokenized base on the tones of the syllables. Digital Signal Processing (DSP) is the computer analogy of the dynamically controlled the ariculatory muscles and the vibratory frequency of the vocal so that the output signal matches the input requirements. It consist of the speech Processing and Sound Processing. The speech processing lookup for syllables matching them to strings (concatenate) and smoothening them while Sound Processing (Speech Signal) process the audio which makes the pronunciation sound audibly. *3.1 Architectural framework for Yoruba Text to speech*

For this research, an architectural framework for Yoruba TTS is developed and presented in Figure 2. The framework defines the components of the system together with the interaction between each component. During the design stage, the architecture of the system was developed taking into account the constraints impose by the user's requirements and the available technology.

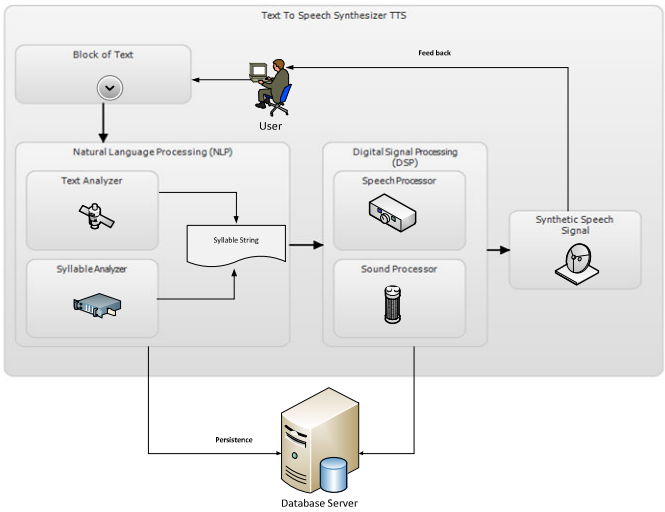

Figure 2. Architectural framework for Yoruba Text to Speech

3.1.1 *Syllable Identification*

NLP will be perform by synthesizer, it will brake every block of text to syllable and identify the Vowel and the consonant vowel (V ,CV) and recognises the tone bearing vowel, it also recognised ẹ ọ, ṣ consonants and differentiate them from the similar consonant. This process serves as normalization of the block of text. The flow chart for this is shown in Figure 3.

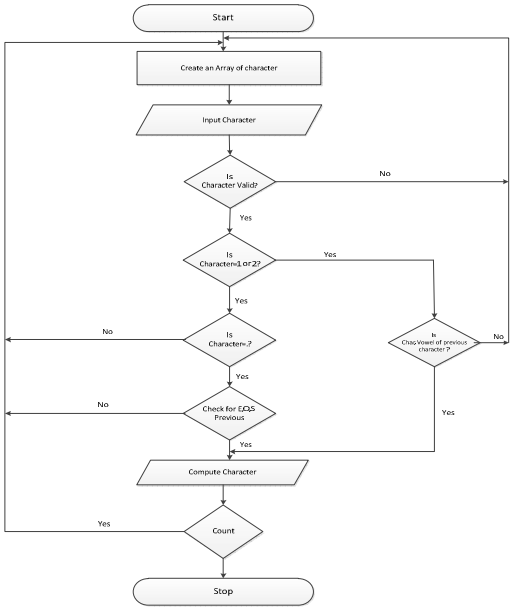

Figure3 Syllable Identification flowchart

3.1.2 Speech Pronunciation

The DSP aspect of the model performs this function, A sound file was created for the selected syllables in three tonne of Yoruba language: High, Mid and Low tones, the recorded syllable sound was matched with the block of supplied text so as to have correct pronunciation. Shown below

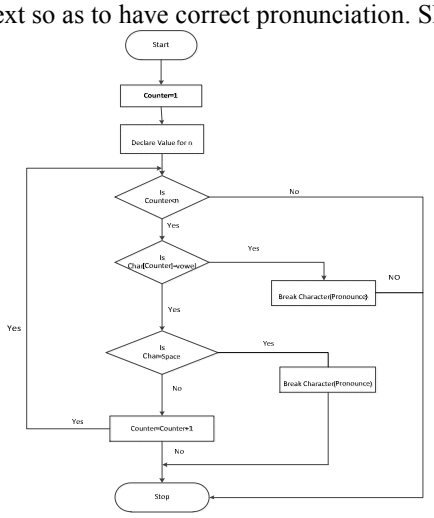

Figure4 Speech pronunciation flow chart

The overview of Application Framework and the sequence diagram of the Developed Yoruba TTS were shown in the figures 5 and 6.

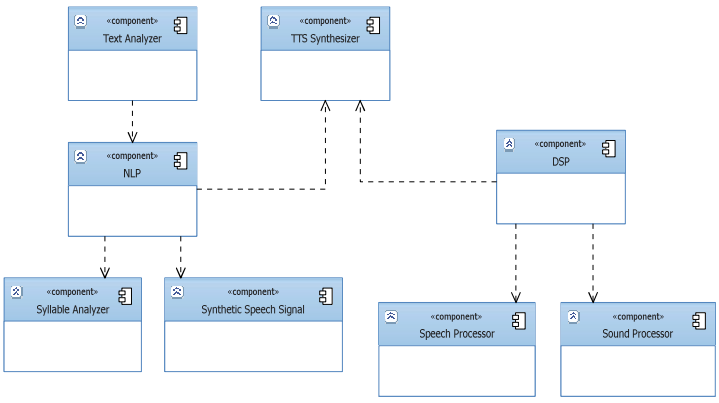

Figure5 Overview of Application Framework

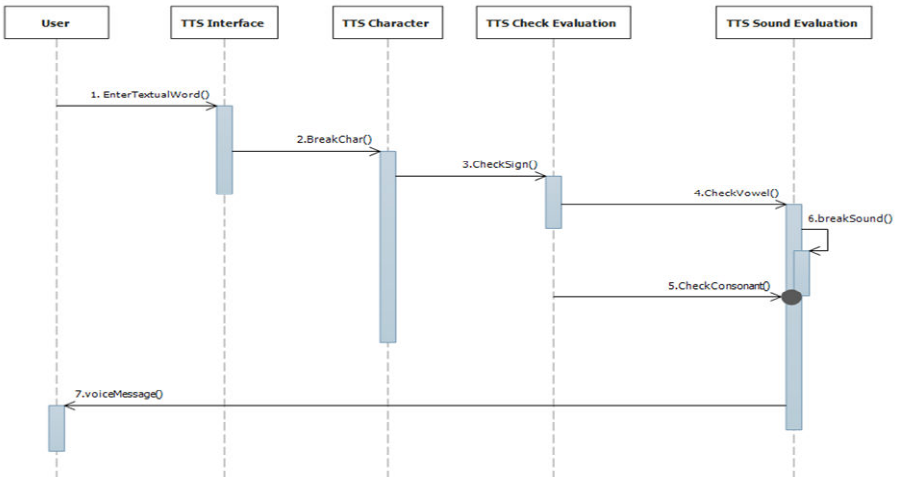

Figure6Sequence Diagram of the Developed Yoruba TTS System

3.2 Implementation Tools chart

The tones of Yoruba syllable were recorded by using a Personal Computer enhanced with voice recording facilities such as head phone with Condenser Microphone (SONY ECM-44S Electrets) by a mastery native speaker of the language in a quiet environment, the recorded data were edited using sound editing software (Adobe SoundBooth) where noise and other unwanted pitch were removed. A dialect called 'Yoruba Oyo' was chosen for the recording, this dialect is popular in northern part of Oyo State, South West Nigeria and it is also known as standard Yoruba. Integrating the recorded syllable sound to match the block of supplied text so as to have correct pronunciation was achieved with C# programming language and was implemented on Microsoft Visual Basic environment.

3.3 Performance Evaluation of the Developed System

The developed system was evaluated to determine its relevancy based on users' assessment in terms of ease of usage, acceptability and naturalness. When responding to a Likert questionnaire items, respondents specify their level of agreement to the statement. The most common scale is 1 to 5. The results and statistical analysis of the evaluation are presented in the table 5 and 6. The likert items (i.e. each question asked in the questionnaire) provide a metrics on which the performance parameters mentioned above were formulated. Survey targets were set for each evaluation parameter of the developed system.

4. Result and Discussion

Figure 6 illustrates the main user interface, it was designed in such a way that user would easily understand the procedure of operation, the following tools display on the interface:

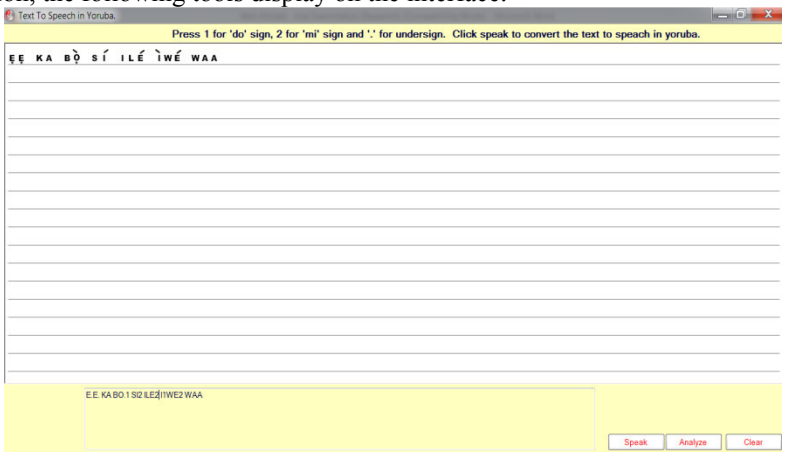

Figure6 Sample Screen of Yoruba TTS System

4.1 Concatenative Synthesis of Yoruba Syllables

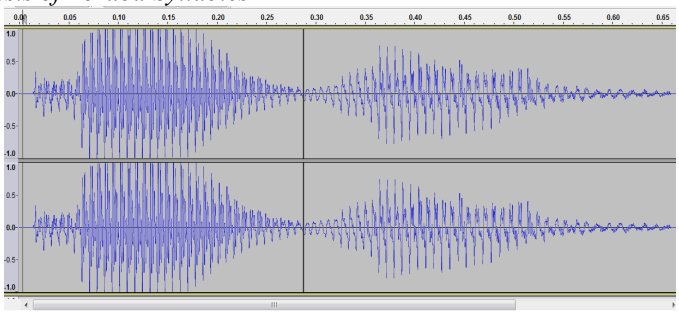

Figure7. wave diagram showing how the word 'Ọwọ1' (Hand) was concatenated

Based on concatenation in figure 7 (stringing together) of segments of recorded syllables, stringing of syllables occurred when the text being supplied was linked to its corresponding syllables in the sound database. The pronunciation of the word 'ỌWỌ' (hand) was achieved by concatenated the two syllables involved (Ọ and WỌ) with the correct tone on each syllable, the sound of each syllable were fetched from the sound database and then stringed them together which then produced distinct sound.

The first column of each table represents the question number of an item on questionnaire while data in the split cells represent the frequency of the response in number. The response mean and the response mode were also indicated. Furthermore, the bar chart of the numeric frequency of the respondents is shown on Figure 7 and 8. Summarily, with this result the naturalness and intelligibility of the developed system can be rated base on perception of the randomly selected users with the likert item in Table 5. It can be seen that the performance, voice, and ease of usage of the developed system were rated averagely okay by 70 % of total respondents. Table 5 shows 35 out of 50 respondents agreed that the developed system had ability to produce sound like Yoruba natural speaker tone, also the frequency shows that between 35 respondents believed that this TTS can be used as a teaching aid and also for aiding learning for visually challenge person.

Table5.Data Analysis of the Administered Questionnaire

QUE	Likert Item	Very High	High	Medium	Low	Very Low	Weighted Score	Rank
X3	How would you rate the performance of the developed TTS system?	2	15	32			3.39	3
X4	How would you rate the ease of use of this application?		9	36	5		3.08	3
X5	How would you rate the voice quality produced by the developed system?	3	13	29	5		3.28	3
X6	How would you rate the tonnation of the Yoruba alphabets display on the screen?	1	9	31	5	4	2.96	3
X12	What is your rating of the naturalness of the output from the developed TTS?	2	10	31	5	2	3.1	3
X13	What is your overall performance assessment of the developed TTS?	3	20	27			3.52	4

Table6.Data Analysis of the Administered Questionnaire

QUE	Likert Item	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	Weighted Score	Rank
X6	Would you agree that the sound produce by the developed TTS was close to Natural human voice?	1	9	5	31	4	2.44	2
X7	Would you agree that the response time i.e. time duration for producing some was of no significance?	15	11	19	4	1	3.68	4
X8	Would you agree that the developed TTS system is reliable to convert Yoruba text to spoken expression?	3	4	5	34	4	2.36	2
X9	Would you agree that the developed Yoruba TTS can serve as a teaching aid for Yoruba language?	5	4	7	32	2	2.56	3
X10	Would you agree that the developed system will be of help to visually challenged person?	1	12	1	34	2	2.52	3

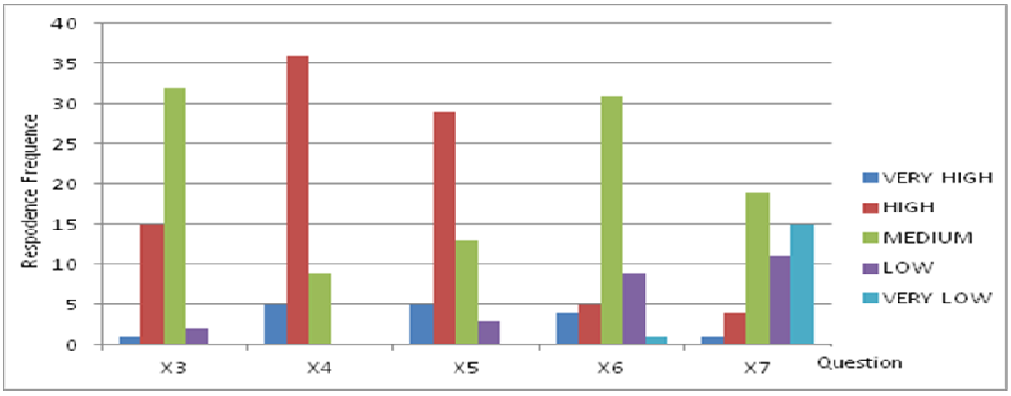

Figure7. Bar chart representation of Numeric Frequency of Response of Table 5

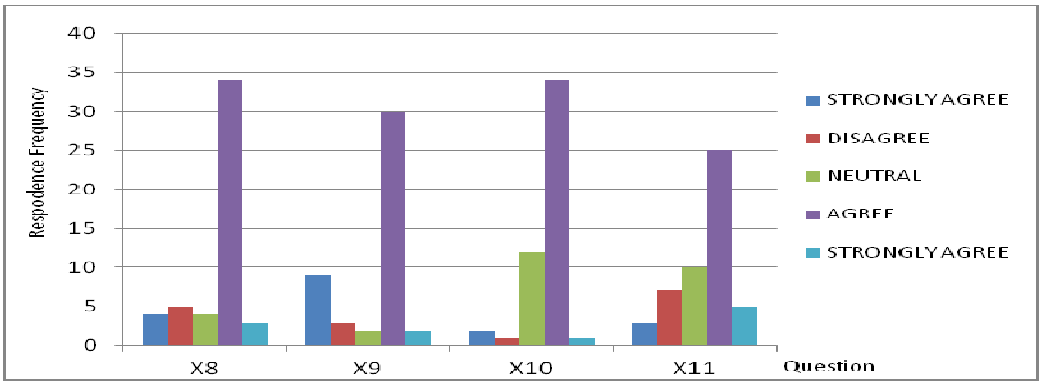

Figure8. Bar chart representation of Numeric Frequency of Response of Table 6

5. Conclusion

A Yoruba TTS system has been developed and tested. Its performance was found to be acceptable, the developed TTS system would enhance learning of Yoruba Language also Consonant-vowel sound syllables database and alphabet database of the developed system could be adapted for any relevant application in Yoruba language development. Although the pronunciation was not actually smoothened but the system still provides the user with the capabilities of Yoruba word pronunciation.

References

Odejobi O.A., Beaumont A.J. and Wong. (2006): *intonation contour realisation for Standard Yoruba text to speech synthesis A fuzzy computational approach* computer speech and Language, Vol. 20, pp 563-588.

Odetunji O. A., (2006): "*A Quantitative Model of Yorùbá Speech Intonation Using Stem-ML* A selected Proceedings from Conference on Human Language Technology for Development, Alexandria, Egypt.

Dutoit T. and Leich H., (1993): Text-*to-speech synthesis based on a MBE. Re-synthesis of Segements Database* Speech communication Journal vol.13 No 15 pp. 435-440.

Every Culture (2012): Countries and their Cultures. Consulted on the 26th June 2012. www.every culture.com

Alan W., (2002): "Perfect synthesis for all the people all of the time". Keynote address at the IEEE Workshop on Text to Speech on 30th of September. vol., 13 no7 pp 22-30accessed on 12th of June 2011 at: http://www.cs.cmu.edu/-

Van P.H., Richard W.S., Joseph O. and Julia H., (1997): "Processes in Speech Synthesis" Springer Press: ISBN 0387947019

Mogey N., (1999): "So you want to use a Likert Scale?" Learning Technology Dissemination initiative, Heriot-Watt University.

Olanike O.,Ola Orie., (2006): "Acquisition and Yoruba Tones: issue and challenges" Selected Proceedings of the 36th Annual Conference on African Linguistics. Published by Cascadilla Proceedings Project, Somerville, USA. pp 121-128.

Afolabi A.,(2012): "Development of Yoruba Text To Speech System" M.Tech (Computer Science) Thesis, Ladoke Akintola University of Technology, Ogbomoso, Nigeria.