Vol 8. No. 1 – March, 2015

African Journal of Computing & ICT

www.ajocict.net

Development and integration of Text to Speech Usability Interface for Visually Impaired Users in Yoruba language.

O.O. Adeyemo & A. Idowu
Department of Computer Science
University of Ibadan
Ibadan, Nigeria.
E-mail: wumiglory@yahoo.com

ABSTRACT

Text to Speech system (TTS) cannot be overlooked because of the important roles it performs in enabling the user to access the voice output of the input text. Text To Speech is a system that takes text as input and produces the equivalent natural human voice of the text input. The system has been applied in different native languages outside Nigeria. In this paper, we considered development of Text to speech in Yoruba language to assist Yoruba language speaking people especially the visually impaired users, it also helps the users that want to learn Yoruba language from scratch, the aspect of how to learn, pronounce Yoruba language syllable formation from consonant and vowel. The performance of the model was also measured using a standard error metric. There are several methods which TTS system have been applied in other part of the world but concatenative method of speech synthesis through syllable construction algorithm; was implemented using C# Programming language. The performance of the system was measured and the quality of synthesized speech was assessed using Mean Opinion Score (MOS) tool and the system is found to be 4.46 and 3.82 respectively. The MOS scale interpreted the system as a good system.

Keywords: HCI, Text-to-Speech, grapheme, phoneme, waveform.

African Journal of Computing & ICT Reference Format:
O.O. Adeyemo & A. Idowu (2015). Development and integration of Text to Speech Usability Interface for Visually Impaired Users in Yoruba language. Afr J. of Comp & ICTs. Vol 8, No. 1. Pp87-.94.

I. INTRODUCTION

Human Computer Interaction (HCI) is one of the new research areas in Computer Science, a discipline concerned with the study, design, construction and implementation of human-centric interactive system, HCI assistive and user-centric attributes is the reason for writing this paper [3]. Speech technology together with computational resources has well advanced in various dimensions. This is as a result of shifting in traditional techniques of assessing various computing functionality from the use of mouse and keyboard as a means of input for computing to the use of speech technology, especially to assist the visually impaired.

On many occasions, the source of information originates from a human being, and is ultimately to be used by human being [7]. There is thus a need for man-machines interaction, in both directions, which will be effective. A convenient way in many cases is in the form of speech, because speech is the common and most widely used mode of communication between human beings. In general, speech system can be categorized into two broad categories: Speech synthesis and Speech recognition. *Speech synthesis* is the process of generating spoken language succession from arbitrary text while *Voice recognition* is the process of converting spoken language into computer understandable text or translate it as command [13].

Some of the speech technology applications include the following; *reading machine, voice interface, pocket translator, talking word processor, auto attendants, speech transcription system, voice dictionary, travel reservation agent, call centre automation* and so on. The goal of the new trend is to make machines speak and hear like humans do. With this advancement, the world of human communications would be substantially expanded. Achieving the goal for local languages like Yoruba language would also provide novel knowledge about cognition, understanding, and mechanism of speech production and hearing, certainly useful for various aspects of human life.

1.1 Basics of Speech Synthesis

Speech synthesis gives us the ability to convert arbitrary text to an audible audio and natural sound format where the ultimate importance is to convey textual information to the people in natural voice. The major purposes of speech synthesis techniques is to convert a chain of phonetic symbols into artificial speech, to transform a given linguistic representation and to generate speech automatically with information about intonation and stress i.e. prosody. TTS system contains two components: they are Natural Language Processing (NLP) and the Digital Signal Processing (DSP) components [6].

Natural Language Processing (NLP) is targeted to produce phonetic transcription of the text, together with the desired intonation and rhythm. The Digital Signal processing (DSP) transforms the symbolic information it receives from the NLP module into speech [4]. With the help of these two components, TTS systems involve the following stages in the process of converting written text into speech. *These steps are text analysis, phonetic analysis and prosodic analysis, and speech generation, they are explained below:*

Text Analysis: text analysis involves breaking down of raw text into pronounceable words. It involves the work on the real text, where many Non-Standard Word (NSW) representations appear. For example, the text may contain numbers (year, time, ordinal, cardinal and, floating point), abbreviations, acronyms, currency, dates, URLs. All of these non-standard representations should be normalized, or, in other words, converted to standard words.

Phonetic Analysis: It is simply conversion of analyzed token into pronounceable chunk. The phonetic analysis module takes the normalized word strings from the text processing module and produces a pronunciation for each word. The pronunciation is provided as a list of phones, a syllabic structure and lexical stress. The method for finding the pronunciation of a word is either by a lexicon or by letter to sound rules.

Prosodic Analysis: This stage is where certain properties of the speech signal such as pitch, loudness and syllable length are processed. Finding correct intonation, stress, and duration from written text can be challenging, prosodic features segment speech chain into groups of syllables. This gives rise to the grouping of syllables and words in larger chunks [6].

1.2 Methods of Text to Speech System

A good speech synthesis system must exhibit intelligence and naturalness which forms major characteristics of an ideal synthesizer. The choice or the method employed depends on the language, system and the platform used. There are various ways by which speech synthesis can be carried out. Three major methods are very important and discussed in this paper. They are;

- Articulatory synthesis
- Formant synthesis
- Concatenation synthesis

Articulatory-speech-synthesis: Articulatory based speech synthesis technique attempts to parameterize the human speech production system directly, that is, it tries to model the human vocal organs as perfectly as possible in such a way that each synthetic speech will be similar to the natural speech produced by each vocal organs. This technique basically uses five articulatory parameters: area of lip opening, constriction formed by the tongue blade, opening to the nasal cavities, average glottal area, and rate of active expansion or constriction of the vocal tract. Experiments with articulatory synthesis systems have not been as successful as with other synthesis systems but in theory it has the best potential for high-quality synthetic speech. [8].

Formant Speech Synthesis: Formant synthesizer uses a simple model of speech production and a set of rules to generate speech. While these systems can achieve high intelligibility, their naturalness is typically low, since it is very difficult to accurately describe the process of speech generation in a set of rules. Formant synthesizers may sound smoother than concatenation synthesizers because they do not suffer from the distortion encountered at the concatenation point as human speech sample is not used at runtime. To reduce this distortion concatenation synthesizers select their units from carrier sentences or monotone speech. The synthesis thus consists of the artificial reconstruction of the formant characteristics to be produced [14].

Concatenative Speech Synthesis:
Concatenative synthesis is process of stringing together segments of recorded speech. It is the so called cut and paste synthesis because short segments of speech are selected from a pre-recorded database and joined one after another to produce the desired utterances. In theory, the use of real speech as the basis of synthetic speech brings about the potential for very high quality, but in practice there are serious limitations, mainly due to the memory capacity required by such a system [15]. The longer the selected speech units are, the fewer problematic concatenation points will occur in the synthetic speech. However, the limitation in concatenative synthesis is the need for more memory requirements as the speech unit increases [10]. Other sub categories of concatenative speech synthesis include *unit selection*, *diphone synthesis* and *domain-specific synthesis.*

2. RELATED WORKS

Noriko Umeda et al 1968 [9] built the first general English text to speech, although the first computer based speech synthesis systems were created in the late 1950s. Noriko and his colleague carried out this work at Electro-technical Laboratory, Japan. Their work was specifically based on English language. Breslow, et al. 1982 described the Texas Instruments *Speak 'n Spell* toy, released in the late 70s, was one of the early examples of mass production of speech synthesis. The quality was poor, by modern standards, but at the time of creation, it was very impressive [2]. Speech was basically encoded using LPC (linear Predictive Coding) and mostly used isolated words and letters though there were also a few phrases formed by concatenation. Simple text-to-speech (TTS) engines based on specialized chips became popular on home computers such as the BBC Micro in the UK and the Apple.).

Odetunji 2008 developed a neural network model using Multilayer Perceptron (MLP) and Recurrent Neural Network (RNN) [11]. The use of neural network knowledge helped him to develop model that could recognize Standard Yoruba tone, having studied the tonal characteristics of Yoruba. The model used fundamental frequency f0 profile of standard Yoruba syllables to distinguish the high, mid and low tone level in Yoruba language. Base on linguistic knowledge, the tonal parameters were selected carefully alongside with acoustic data. He concluded that standard Yoruba tone realization problem could be solved with MLP and RNN and that mid tone has highest accuracy, came out with performance result of 71% to 76%.

Akin afolabi, et al. 2013, developed aText – speech system for Yoruba language. Their design also shows the rate of acceptability of the TTS interface from the population of users captured for the experiment [1].

2.1 Description of Yoruba Standard Word

Yorubá is one of the three major languages in Nigeria. The population of people speaking Yoruba language covers Southwestern Nigeria. It is widely spoken language because of its prevalence both in Nigeria and outside Nigeria like Republic of Benin, Togo and many others [11].

2.2 Yoruba Spelling and Pronunciation

Tone languages, such as Yorùbá and some others are different from languages that have no tone, example is found in languages like English and French [11]. In non-tone language, lexical items are distinguished by the stress pattern on the syllables that constitute an utterance. For example, the English words '***re***cord' (verb) and 're***cord'*** (noun) differ in syntactic class and meaning because of the stress pattern on their component syllables. In the verb '***re***cord' the first syllable is stressed [17]. In the noun 're***cord'*** the second syllable is stressed. In tone languages, tone is used to distinguish lexical items rather than stress. Yoruba has three distinct tones which are HIGH (do), MID (re) and LOW (mi) that can be used to distinguish syllables. The tones are associated with the individual syllables in an utterance. For example, in Yoruba: *jẹ´* (H) [to become], *jẹ* (M) [to eat], *jẹ̀* (L) [to receive lash] differ in meaning because of the tone associated with each syllable.

2.3 The Yoruba Alphabet

The Standard Yoruba alphabet has 25 letters which is made up of 18 consonants (represented by the graphemes: (*b, d, f, g, gb, h, j, k, l, m, n, p, r, s, s., t, w, y*) and seven vowels (*a, e, ẹ., i, o, ọ, u*) while the Latin Letters 〈c〉, 〈q〉, 〈v〉, 〈x〉, 〈z〉 are not used. There is also inclusion of a diagraph 〈gb〉 which contains combination of two consonants together that form a unit.

Yoruba Vowel

Phoneme	Orthography	Examples	English
/a/	a	ajá	'dog'
	àbá	'motion'	
/e/	e	ewé	'leaf'
		èté	'lips'
		as in English	bait
/ ẹ /	ẹ	ẹ̀ j ẹ̀	blood
		ẹ̀ f ẹ̀	jest
		as in English	'bet'
/i/	i	ìrì	'dew'
		ìdí	
'buttocks'			
		as in English	'beat'
/o/	o	owó	'money'
		òdo	'zero'
		as in English	'boat'
/ ọ / or /⊃/	c	ọfọ̀	
incantation			
		ọjọ́	day
		as in English	
'bought'			
/u/	u	ojú	
'eye/face'			
		òwú	'thread'
		as in English	'boot'

Adapted from:
African Studies Institute manual, University of Georgia, USA

3. METHODOLOGY

This section defines various activities and steps undertaken in the process of designing a befitting synthesizer, in as much as analysis and design frameworks provide a solid foundation for any application. Descriptions of the Yoruba language text to speech system are therefore analyzed. And we make use of concatenative method of speech synthesis. That is, linking together of pre-recorded Yoruba syllable stored in a database.

3.1 General Structure of TTS System

Text-to-Speech synthesis is an attempt to artificially produce human speech. A computer system used for this purpose is called a speech engine.
TTS system transforms any text into speech in real time. It literally reads out loud any written information with a smooth and natural sounding voice.

The structure comprises a *Natural Language Processing module* (NLP) and a *Digital Signal Processing module* (DSP). NLP is capable of producing a phonetic transcription of the text read, together with the desired intonation. And whatever processing done by the Natural Language Processing unit of the system is received by the Digital Signal Processing (DSP) for transformation into speech wave form. The functional diagram of general Text to Speech is depicted below:

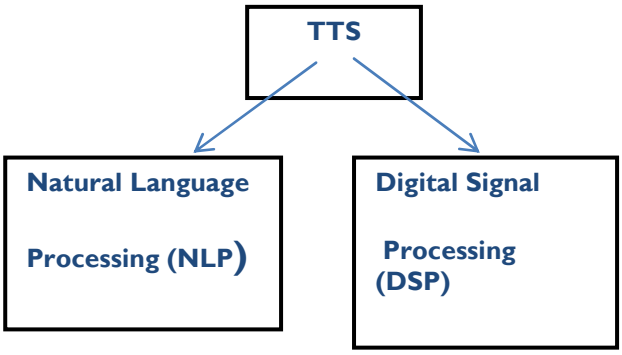

Figure 1: general structure of text-to-speech system.

Yoruba Syllable Construction

Syllable is defined to be a unit of pronunciation that has one vowel sound with or without surround consonant which form all or part of a word. The most common sequence is that of a vowel preceded by a consonant e.g. 'wa' meaning "come" which has the most usage in word construction. A word can start either with a vowel or a consonant. The three possible syllable structures of Yoruba are Consonant+Vowel (CV), Vowel alone (V), and syllabic Nasal (N) which means there is possibility formation of

- CV – consonant vowel e.g "wa" (come), "pa" (kill)
- V – vowel e.g. "a" (we), "o" (she/he – third person singular pronoun)
- N – Nasal e.g. an, un – work in combination with other consonants

A typical Yoruba must start with either vowel or consonant, and can be ended with vowel, example is "Aja" meaning (dog), "baba" meaning (father). But consonant cannot end Standard Yoruba words, for instance, saying "babaj", "kolat" – are invalid words in Standard Yoruba language, every word must end with vowel or nasal sound only. The syllable is intended to deduce all the possible combination of syllables to enable a vast variety of words to be catered for. The reason for this syllable construction is that, if forms basic speech unit for concatenation

1) Analysis of The Proposed System

Synthesizer Design of the Proposed System.

The task of speech synthesis is to map a text to a waveform. Figure 2 shows the main components designed of Yoruba speech synthesizer. It mainly constitutes the natural language processing module, the digital signal processing module and a Yoruba syllable database which augments the speech synthesis process. Given an input text, the Yoruba synthesizer performs the necessary text analysis. This includes;

- Breaking down of the text input into separate words through blank spaces and full stops.
- Tokenization process takes place which forms an integral part in the synthesizers' design which basically normalizes numbers.
- The phonetic analysis component performs grapheme to phoneme conversion, which extracts individual phonemes making up the whole word and passes them to the prosodic analysis featuring the intonation, pitch and duration attachment.

Architectural Framework of The New System

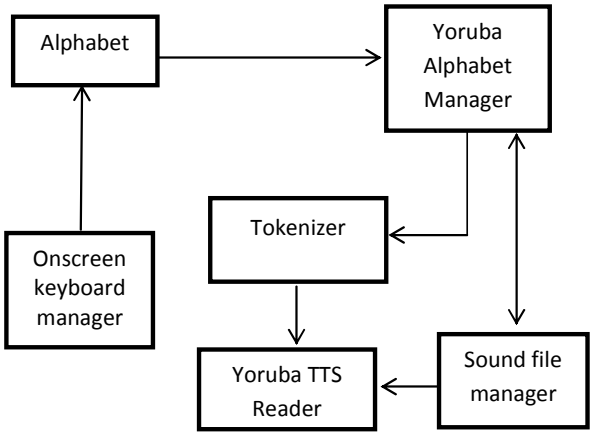

Figure 2: Architectural framework of the new system

The structure of the new system is broken down into different modules that are design to carry out several functions. The arrows point to the module that is used by another module in its operation, for example in the diagram above,

Description of the New System
the YorubaAlphabetManager makes use of the Alphabet module, likewise it (YorubaAlphabetManager) must check from the SoundFileManager to check if there exist such file in the sound file which it can use to concatenate syllables together.

YorubaAlphabetManager – is responsible for the processing and production of alphabet input stream from the backend, constructed from the Alphabet module.

OnscreenKeyboardManager – it is a graphic user interface on the new system that helps the user to write strings of characters.

Tokenizer – this module splits the input text into token (small chunks), which can be inform of syllable or vowel.
SoundFileManager – contains pre-recorded syllables, vowel and consonants from where any matched sound file can be read.

YorubaTTSReader – this is the module that collects tokens from the tokenizer, match it up with equivalent syllable sounds stored in the SoundFileManager module, if the tokens are the same as the sound, YorubaTTSReader reads the text aloud.

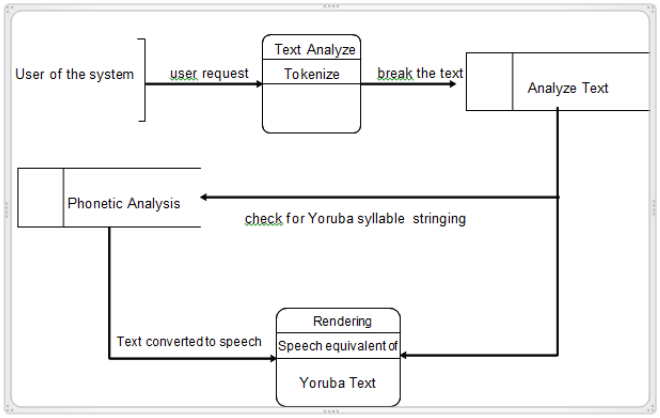

Figure 3: Data flow diagram (Dfd) of the new system.

The algorithm for splitting Yoruba words into syllable.

Start
Read in the text
Break text into characters
foreach of the char in the char array
Get the last char
If the last char is consonant and <>n or m Then
return error
Else
Continue

Get the alphabet accordingly and Create a check point
If the alphabet is consonant then
Continue
Else
Get all the alphabets from the checkpoint to the current index as a syllable
Convert the characters
end the loop
stop the program.

4. IMPLEMENTATION

In this paper, text to speech synthesizer for Yoruba language was designed and implemented. It consists of both coding and integration of class modules which are combined together to form a whole and complete system. The implementation was done through transformation of the design using C-sharp programming language for the implementation. There is therefore need to test the system, steps are taken to check the extent to which the synthesizer works, based on what to report, provides findings and suggest further research areas.

4.1 Multimedia Kit Used in Studio

The recording was done under good acoustic condition at Master Sound Studio. The voice recorded was a female voice.

We use female voice so as to have a high pitch and clear sound in the production of the text to speech system. The recording though was not easy and stressful, took up to two weeks before we can get appropriate pronunciations for all the vowels and syllables, just a single syllable can be re-recorded for several times if it does not pronounce the desired or expected sound. In the recording of the syllable, each consonant is matched with every of the vowel sounds and are recorded at length, Cubase voice station software was later used to edit and cut sounds into syllable segments.

4.2 Preparation of Voice Database via Sound Studio Recording

Building the sound file for Yoruba synthesizer is an essential aspect when building a text to speech system for Yoruba language, because from this sound database where sound will be produced to match various texts input, eventually uttered by digital signal processing component of the synthesizer.
To prepare Yoruba sound file, we first created inventory of syllable pronounceable in Yoruba.

Yoruba language has 7 vowels and 18 consonants. There are 126 syllables that can be derived for a single tonal level. The three tonal level in Yoruba are “do” (high), “re”(mid), “mi”(low). If we form syllables for the three different tonal levels, we will have 126 x 3 = 378 syllables. Examples of those syllable are “ba”, “be”, “be,”, “bi”, “bo”, “bo,”, “bu”. Nasal vowels (an, en, e,n, in, on, o,n, un) were also part of the inventory created. All of them were recorded. The appropriate set of folders and files were created through recording with Cubase voice station software. All the files are stored in a .wav form with a standard approach in naming them to avoid later confusion. Speech signals are recorded by a close talking microphone using a sampling rate of 16 kHz

Ba	be	be,	bi	bo	bo,	bu
da	de	de,	di	do	do,	du
fa	fe	fe,	fi	fo	fo,	fu
ga	ge	ge,	gi	go	go,	gu
gba	gbe	gbe,	gbi	gbo	gbo,	gbu

Figure 4: few of the Yoruba syllables recordings

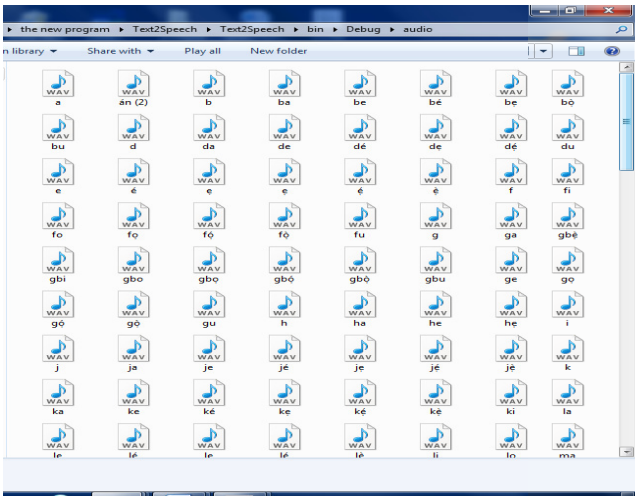

Figure 5: Yoruba Database Recordings

Main Features of the New System

- ✓ **Welcome Interface/Home page:** it shows first point of contact with the system.
- ✓

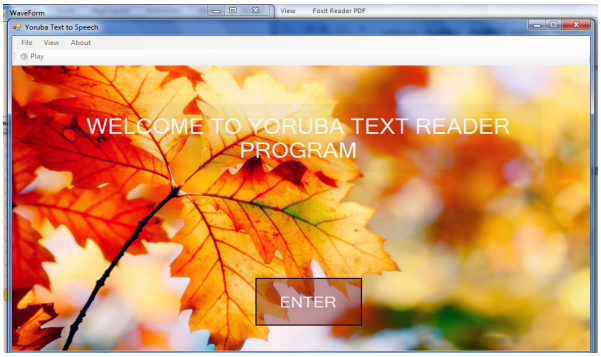

Figure 6: Welcome interface of the system

- ✓ **Main page:** The main page has features that enable user to load text file which can be read. The main page is initially disabled until text file is loaded into the main page which makes the main page to be enabled; the saved Yoruba text is displayed on the main page to be read.

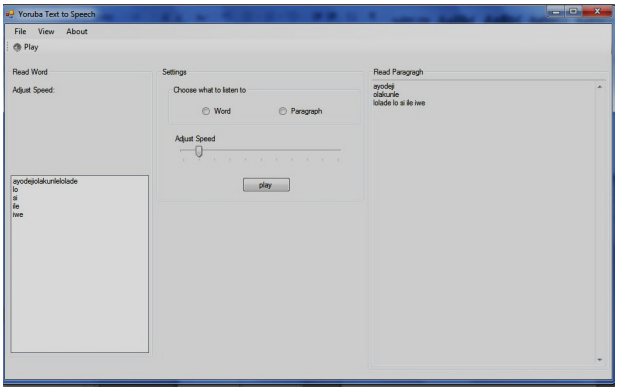

Figure 7: main page of the system

- ✓ **"Word Read" page:** is the page that allows user to read the Yoruba text at word level, that is, Yoruba word can be read, one after the other.

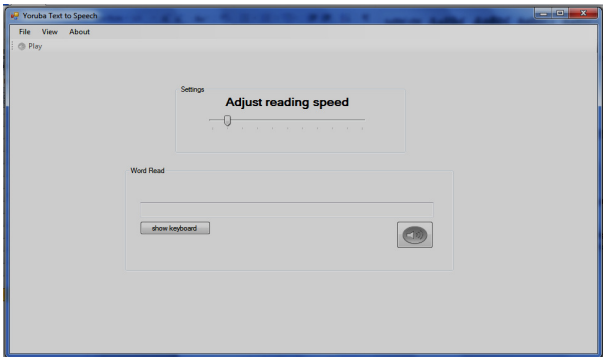

Figure 8: "Word Read" page of the system

Evaluation and Result of the New System

The performance assessment of the new system is described in this section; we carefully notice and observe the consistency of the system in term of its intelligibility and naturalness.

Intelligibility – describes how accurate or correctly the system pronounces or how closer is the pronunciation to the actual word.

Naturalness – describes whether the system produces voice as natural as that of human voice.

The first experiment is on the performance of the system that is assessed on word level. The test consists of 61 Yoruba words selected through the help of a domain expert. We listened to one word played from the synthesizer at a time and marks on the answer sheet for the synthesized word, to know which one is correctly pronounced and the ones not correctly pronounced.

Table below shows the analysis made on the performance evaluation of the test dataset.

Table 1: Performance Measure of the Yoruba TTS.

.

Performance Measure on the Test Dataset				
	Correctly Pronounced	Partially Pronounced	Not Correctly Pronounced	Total
No. of Words	38	11	12	61
% of Words	62.3%	18.0%	19.7%	100%

The overall performance of the system is measured in terms of total number of correctly pronounced words over the total number of words played i.e. (correctly pronounced words/total numbers of word x 100%). Finally by calculating the number of words which are correctly pronounced, the performance of the system is found to be 62.3%. As observed from the analysis, words that are not found in the compiled sound file are not properly pronounced by the system (nasal vowels are not taken care of). This causes a little degradation in the performance of the system.

Second performance evaluation focuses on the *intelligibility* and the *naturalness* of the new system. Mean Opinion Score (MOS) evaluation technique was used to evaluate the synthesized Yoruba Speech. Mean Opinion Score (MOS) is an evaluation technique which provides a numerical measure of the quality of human speech. The scheme uses subjective tests (opinionated scores) that are mathematically averaged to obtain a quantitative indicator of the system performance. To determine MOS, a number of listeners rate the quality of test sentences read aloud from the new system female speakers [16]. A listener gives each sentence a rating as follows:

Table 2: Scales used in Mean Opinion Score (MOS) Technique

Value	MOS
5	Excellent
4	Very Good
3	Good
2	Fair
1	Bad

In the second evaluation, fifteen native speakers of the Yoruba language were invited. Then the evaluators provide their ranks based on the MOS scale. In both cases a questionnaire is used to collect the evaluator's opinion. To evaluate the synthesizer's intelligibility and naturalness five sentences in Yoruba are prepared as a test data for Yoruba synthesizer.

All words used in the sentence are found in the normal Yoruba Literature. Then the selected individuals listen to the synthesized waveform from the synthesizer and evaluate naturalness and intelligibility based on the MOS scale.

Table 3: Overall average score of Yoruba Text to Speech

	1st Sentence	2nd Sentence	3rd Sentence	4th Sentence	5th Sentence	Average Score
Naturalness	4.7	4.5	4	4.5	4.6	4.46
Intelligibility	3.8	3.7	3.6	3.9	4.1	3.82

The resultant naturalness of the new system using five Yoruba sentences as test data yielded 4.46. This shows that the synthesizer is "very good" according to Mean Opinion Score (MOS) Scale. While the overall intelligibility of the new system is found to be 3.82, which is interpreted on the MOS scale as "good" approaching to "good" if approximated. These values look encouraging and show that a more complex and better system of this kind is achievable.

5. CONCLUSION & CONTRIBUTIONS

By using these Speech Synthesis ideas, we were able to develop a system that pronounces common valid Yoruba words provided by a user and extended them to include sentences. Clearly a Yoruba Text-to- Speech System is a viable and an achievable task. Yoruba Text To Speech was tested and evaluated using Mean Opinion Score experimentation. During the experiment, a workable and performance effective system with 62.3% result was obtained; 3.82 MOS score in intelligibility and 4.46 MOS score for naturalness, which shows that the system is good. Also, Yoruba speech corpus was prepared from scratch in studio which made the system to give quality sounds. Finally the developed model improves interface for the visually impaired, learning system for Yoruba language, performance measure carried out to ensure reliability of the model.

6. FUTURE WORKS

Further work should be carried out to study issues on speech synthesis and provide algorithms for prosody to achieve proper synthesis. Likewise the system can be extended to handle non standard words (NSW) like numbers and dates by creating a parse algorithm.

REFERENCES

[1] Akin afolabi, et al. (2013). Development of Text to Speech System for Yoruba Language. International Conference on Engineering and Technology Research. Vol.4, No.9, 2013 Special Issue.
[2] Breslow, et al.(1982). United States Patent 4326710: "Talking electronic game" April 27, 1982
[3] Card, et al. (1980). "The keystroke-level model for user.
[4] Daniel, J., and James, H., (2000). "Speech and Language Processing" Prentice Hall
[5] Degen J., Wien, Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine ("Mechanism of the human speech with description of its speaking machine,"), Germany.
[6] Dutoit, T., (1997). "A Short Introduction to Text-to-Speech" Kluwer Academic Publishers, Dordrecht, Boston, London.
[7] Holmes, W., (2003). "Speech Synthesis and Recognition" Taylor and Francies New Fetter Lane, London ECAP 4EE.
[8] Kleijn B., et al., (1998). "On the Use of Neural Networks in Articulatory Speech Synthesis". Journal of the Acoustical Society of America, JASA Vol. 93(2): PP. 1109-1121.
[9] Klatt, D. (1987). From Text to Speech: The MITalk system. Cambridge University Press. ISBN 0-521-30641-8.
[10] Lemmetty, S., (1999). "Review of Speech Synthesis Technology" Master Thesis, Helsinki University.
[11] Odetunji A, (2008). Recognition of Tones in Yoruba Speech. Obafemi Awolowo University, Ile Ife, Nigeria.
[12] Performance time with interactive systems". Communications of the ACM 23 (7): 396–410. doi:10.1145/358886.358895
[13] Samsudin, N., et al., (2002). "A Simple Malay Speech Synthesizer Using Syllable Concatenation Approach".
[14] Styger, T., and Keller, E., (1994). ,"Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges" Formant Synthesis. In E. Keller (ed.)
[15] Vosnidis, C., (2001). "Speech Synthesis by Word Concatenation" B.Sc. Thesis, University of Crete.
[16] WhatIs.com. Posted by Margaret R., May 2011. "Definition of Mean Opinion Score (MOS)." Retrieved Sept. 2014.
[17] Xu.Y.(2005). Understanding tone from the perspective of production and perception. Language and Lingustics, 5:757–797, 2005.