International Journal of
Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2019.33568

Volume 5, Issue 10
October - 2019

Text to Speech Synthesis System in Yoruba Language

IKANI Lucy Hassana[1] and MUHAMMAD SANUSI [2]

Research Scholar[1], Lecturer[2]

[1-2]Department of Computer Science,

University of Abuja, Abuja,

Nigeria

Abstract

Previous and existing studies convert texts to speech (English Language) without consideration for indigenous languages. This project work is an effort to design a speech extension that can convert the highlighted text displayed by a web browser to speech considering the indigenous Yoruba language through a designed software tool specifically attached to Mozilla internet as a plug-in. This software tool allows the user to highlight the text on the web page with the mouse. The selected text provides an input to the speech program that runs in the back end. The objective of this project work is to design a program that will help people with weak eyesight, those that lack in pronunciation skills and those learning the Yoruba language as an additional language skill; it is also designed to solve problems that exist when using the manual system which include difficulty in reading and misinterpretation of information. The methodology used to convert the text to speech are the API (Google-translate, speech'.org and ttsyoruba.com) and the programming language used for the web is JavaScript. A text to speech extension is an application that converts the highlighted texts into spoken words by analyzing and processing the text in the Yoruba language. The use of the text to speech is to read out the highlighted text to the user in the Yoruba language which can be saved as an audio file.

Key Words: *TTS, Yoruba Language, Web Browser, Speech Synthesis, Concatenative Synthesis, Digital Speech Pronunciation.*

1. INTRODUCTION

Web browsing is gaining popularity with an exponentially growing number of users each year. Information on the internet is used by various set of people; a student who wishes to find solutions to the home work question, a researcher in need of some information on the research work, a businessman wishing to invest in a business, or a general user trying to know about the weathers, sports or some social events.

In the course of this study, the root of the problem associated with the use of the internet for browsing purposes was traced to the interpretation of the web content and the conversion of a plain text to speech in indigenous languages such as Yoruba language. The inability of people with lack of pronunciation skill to interpret and understand the information conveyed by a web browser can lead to multiple problems to do with user understanding.

This research work is centered on or limited to text to speech extension for internet browser. A study conducted on other related research works revealed that various researchers have tried to solve the problem by providing various standalone application software which are time consuming and can only solve a fraction of the problem. One of the most common and familiar examples of such software is the Interactive Voice Response (IVR) system.

The IVR is being used in application such as voice portals where automated voice services can be accessed over the phone and in the non-phone based voice application such as the embedded home appliances and automobiles. The IVR can be used with anything like checking the bank balance, finding the estimated delivery time of a packet or status of the flight departure, it is all done without talking to actual humans.

2. AIM AND OBJECTIVES OF THE STUDY

The aim of this project work is basically an improvement on existing studies towards the design of a text to speech plug-in for Mozilla fire fox browser that converts a webpage text to an indigenous Yoruba language speech.
The objectives of this project are as follows:
1. To design a program that can ease the understanding of a Yoruba language reader and enhance reading skill;
2. To make easy the readability of a webpage in the Yoruba language;
3. To convert a plain webpage text into a Yoruba language speech;
4. To make the browser more user interactive and language convenient;
5. To design an application that will enhance correct interpretation of information contained on a webpage through the Yoruba language.

3. SIGNIFICANCE OF THE STUDY

The project has its relevance drawn from the need to develop and implement a text to speech that would be of immense benefit to Yoruba language learners and users of the internet with limited or no reading skills or bad eyesight. The speech extension for internet browser will help to handle the problem associated with the manual reading of web browser contents. It will serve as an automated text to speech conversion plug-in which will assist people with impaired speech problems by interpreting the content of a web browser.

4. LITERATURE REVIEW

Speech is the vocalized form of human communication. It is based upon the syntactic combination of lexical and names that are drawn from very large vocabularies. Each spoken word is created out of the phonetic combination of a limited set of vowel and consonant speech sound units [1].

Text-To-Speech, also known as Speech Synthesis, is the computer production of human speech. It is the process of generating spoken words by machine from written input. Speech is often based on concatenation of natural speech i.e units that are taken from natural speech put together to form a word or sentence. Concatenative speech synthesis, according to [2], has become very popular in recent years due to its improved sensitivity to unit context.

Rhythm also is an important factor that makes the synthesized speech of a TTS system more natural and understandable. The prosodic structure provides important information for the prosody generation model to produce effects in synthesized speech [3].

Text-To-Speech TTS is still very much at infancy as researchers are working round the clock to have a better algorithm. A TTS system developed through the establishment of corpus-based synthesis unit database that includes nasals, tones, stops and sadhi rules [4], subsystems of the system includes text-input system, text-to-sound convert system, training of basic synthesis units, and the acoustic wave play system. The system has a multiple accent corpus-based database which was developed using combination of basic phonemes of vowels, consonants and tones from MLT (Modern Literal Taiwanese) books. It has limited speech input but uses large database to develop the MLT. A concatenative synthesis and bell lab approach (combination of phonetics and linguistic structure) to speech synthesis relies on designing and creating the acoustic inventory of the language by taking real recorded speech, cutting it into segments and concatenating these segments back together during synthesis [2]. The synthesizer then produces a concatenative system, based on a set of prerecorded acoustic inventory elements that represent all the possible phone-to-phone transitions of the language.

An Arabic system that uses a rule-based hybrid system, which is a combination of formant and concatenative speech techniques reduces the vocabulary independence and can handle all types of input text [5]. The system omits some vowels of the language in use and also does not take intonation into consideration.

The use of concatenative synthesis bypasses most of the problems encountered by articulatory and formant synthesis techniques [4]. Most developed systems make use of very large database that can slow the system down and also require lots of memory space. The issue of incorrect labeling due the large database can also lead to poor quality of the system.

In [6] the system contains front-end which comprised of text analysis and phonetic analysis. The unit selection algorithm is based on Viterbi decoding algorithm of the best-path in the network of the speech units using spectral discontinuity and prosodic mismatch objective cost measures in place of HMM. The back-end is the speech waveform generation based on the harmonic coding of speech. The Harmonic coding enabled the system to compress the unit inventory size by a factor of three. Though, the

DOI: 10.31695/IJASRE.2019.33568

system used transplanted prosody which does not take intonation into consideration, where generated prosody would have been more effective for the same purpose.

[7] Presents techniques for speech-to-text and speech-to-speech automatic summarization. It uses a two-stage summarization method consisting of important sentence extraction and word-based sentence compaction. Sentence and word units which maximize the weighted sum of linguistic likelihood, amount of information, confidence measure, and grammatical likelihood of concatenated units, were extracted from the speech recognition results. For speech-to-speech, sentences, words and between-filler units are investigated as units to be extracted from original speech and concatenated for producing summaries.

The proposed system is a concatenative speech synthesizer and combines real recorded speech sounds. It is based on prerecorded speech inputs which represents Yoruba language exhaustively by using all possible forms of syllable in the language, the syllabic structure is generated using vowels (v) only and consonant + vowels (cv). Each word is recognized if it exists in the library or broken down into syllables

4.1 Review of Related Works

The modern TTS system converts text into 'synthetic speech sound in a two-stage process [2]. 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 [8].

A TTS system is composed of two parts a front-end and a back-end. [8] 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 [8].

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 [9]. High level Synthesis method was used to develop a TTS for Yoruba Language by [10].

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 [11]. The focus of this paper is to concatenate some Yoruba syllables to produce a speech.

4.2 Background on Text To Speech System (TTS)

Text to speech synthesis system dates back to the early days of attempting to develop speaking machines to the recent state of the art text to speech development.
It has been shown that an appreciable improvement has occurred in text to speech system over the years starting from the early conception of the ideas to develop such system. Much as noticeable improvement can be seen in the light of text to speech development, it also invokes the need and desire for further improvements in the quality of synthesized speech.
[1] Identified that the 21st century widespread use of computers opened a new stage in information interchange between the user and the computer. Among other things, an opportunity to input the information to computer through speech and to reproduce in voice text information stored in the computer has been made possible. Nowadays the solution of this problem can be applied in various fields. First of all, it would be of great importance for people with weak eyesight. In the modern world, it is practically impossible to live without an information exchange. The people with weak eyesight face with big problems while receiving the information through reading. [12] Pointed out that a lot of methods are used to solve this problem. For example, the sound version of some books is created. As a result, people with weak eyesight have an opportunity to receive the information by listening. But there can be a case when the sound version of the necessary book couldn't be found.
Therefore, the implementation of the speech technologies for information exchange for users with weak eyesight is of a crucial necessity. In fact, computer synthesis of speech opens a new direction for an information transfer through the computer. For today it is mainly possible through the monitor.

DOI: 10.31695/IJASRE.2019.33568

4.3 Speech Synthesis

Speech is the act of producing voice via variation of the air that is emitted by the articulate system [13]. Whilst, speech synthesizer is the artificial production of human speech where a text-to-speech synthesizer should be able to automatically convert any text into signal carrying linguistic information before it is converted into an acoustic waveform using machine. Major purpose of TTS synthesis System is to transform a given linguistic representation, say a chain of phonetic symbols into artificial, machine generated speech with information on phrasing, intonation and stress by means of an appropriate synthesis method.

The modern TTS system converts text into 'synthetic speech 'sound in a two-stage process [9] the first stages 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 is dram from the domain of Natural Language Processing (NLP) and computational Linguistic [14].

The second stage, called Low Level Synthesis (LLS), takes the linguistic description outputted from the HLS stage input and generates the corresponding speech wave form. The ultimate goal of this stage is to generate speech signal which has as much as possible mimics the acoustic behavior of the speech produced by the native speaker reading the text aloud. There are three methods used to realize the LLS modules, these are Articulatory Synthesis Methods, Formats Synthesis Method and Concatenative Synthesis Method. All these methods will be fully discussed in the next section.

A TTS system is composed of two parts a front-end and a back-end. The front-end has two major tasks. First, it converts raw text containing symbols like numbers and abbreviation 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. The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion [8]. Phonetic transcription and prosody information together make up the symbolic linguistic representation that is output by the front-end. The back-end often referred to as the synthesizer which converts the symbolic linguistic representation into sound.

4.4 Yoruba Phonology

The Yoruba alphabet consists of 25 letters which were derived from Latin characters. The Yoruba language learner is fortunate for two reasons. First, with the exception of a few segments, the writing system closely matches the sounds system of the language. Secondly, with the exception of almost the same set of unique sound, the Yoruba segments in many other languages [8]. Note that consonant "gb" is a diagraph i.e a consonant written in two letters Yoruba alphabet is shown below; it consist of 25 characters.

4.4.1 The Upper Case And Lower Case Representation Of Yoruba Alphabet

Aa Bb Dd Ee Ee Ff Gg GB gb Hh Ii Jj Kk Ll Mm Nn Oo Oọ Pp Rr Ss ṢṢ Tt Uu Ww Yy

4.4.2 Yoruba Consonant (kỌnsónántì) and Vowels (fáwẸlì)

Yoruba consonant is made up of 18 alphabets. This consonant letters were part of the Yoruba alphabet but 18 alphabets made up Yoruba consonant while the remaining letters were Yoruba vowel letters.

Orthography Representation of Yoruba consonant are, Bb Dd Ff Gg GB gb Hh Jj Kk Nn Pp Rr Ss Ṣṣ Tt Ww Yy

In Yorùbá alphabets, there are two types of vowels or fáwẹ̀lì: (a). Plain / normal vowels: a, e, ẹ, i, o, ọ, u (b). Nasal vowels: an, ẹn, in, ọn, un.

5. ANALYSIS OF THE PROPOSED SYSTEM

The proposed system is to design a speech application that can convert the highlighted text and displayed the text by a web browser to speech. It is a software tool that is specifically attached to Mozilla internet browser as a plug-in.
This software tool allows the user to select the text on the web page with the mouse, the highlighted text is provided as an input to the speech program that runs at the ***Back-end.*** The attached speech plug-in enables the speech program to convert the highlighted text into speech. The user can also download the text on a web browser.

DOI: 10.31695/IJASRE.2019.33568

The proposed system would have the following advantages over the existing system:

1. Make easy the readability of a web page
2. To design a program that will help people with lack of pronunciation skills
3. To convert web page text into speech
4. To make the browser more user interactive
5. To design an application that ensure the information contain in a web page is interpreted correctly

5.1 System Algorithm

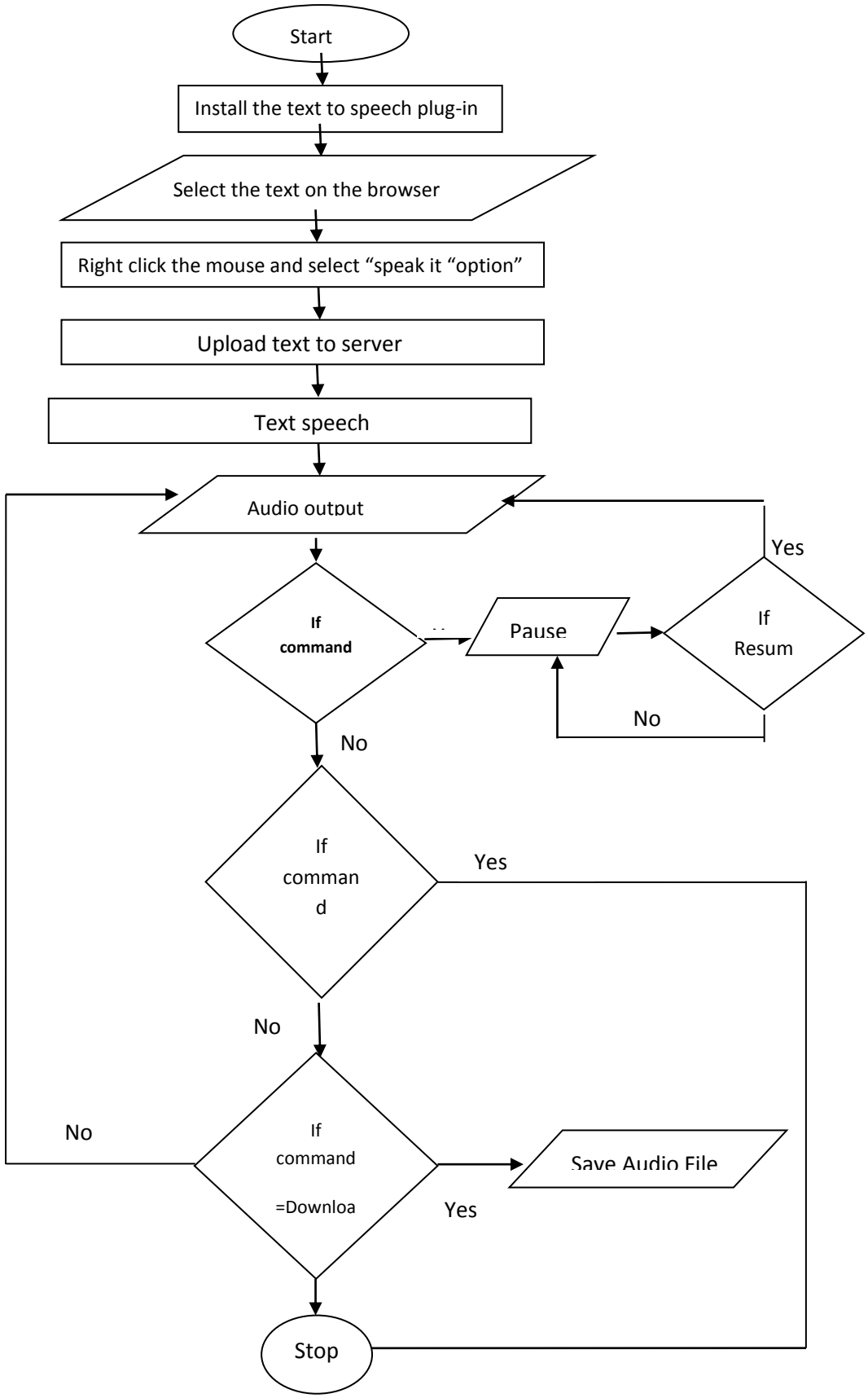
STEP 1: Start
STEP 2: Select text from the browser
STEP 3: Select the convert to speech Option
STEP 4: Check selection text language against user language in settings
STEP 5: If user language in settings = selection text language then
Goto step 5
Else
Convert selection text language to user language based on settings (Using google Translate API)
Goto step 5
End If
STEP 6: Send selection text to the appropriate API and convert to speech
STEP 7: Begin speech output/Reading of the selected text
STEP 8: Selection an option on the speech form
STEP 9: If option is pause then
Pause text reading
Else if option is play/Resume then
Start/Resume text reading
Else if option is stop then
Stop text reading
Else if option is download then
Download speech as mp3
Else if option is close then
Close the speech window
End if
STEP 10: Stop

5.1.1 Option Page Algorithm

STEP 1: Start
STEP 2: Select the tool bar icon or option menu
STEP 3: Select preferred language and voice
STEP 4: Save preference
STEP 5: Stop

DOI: 10.31695/IJASRE.2019.33568

5.2 System Flowchart



5.3 Syllable Identification

The identification process will be performed by synthesizer, it will brake every block of text to syllable and identify the vowel and the consonant vowel (V,CV) and recognizes the tone bearing vowel, it also recognize ẹ, ọ, ṣ consonants and differentiate them from the similar consonant. This process serves as normalization of the block of text.

DOI: 10.31695/IJASRE.2019.33568

5.3.1 Syllable Identification Algorithms

STEP

1: Start
2: Declare Word (n) character as array
3: Supply Character into word
4: Check if character is valid
IF Yes GOTO STEP 5 IF No GO To STEP 3
5: CHECK if Character is 1 or 2
If YES GOTO STEP 6 If NO GOTO STEP 7
6: CHECK IF Character is Vowel or Consonant
If YES GOTO STEP 9 IF NO GOTO STEP 3
7: CHECK if Character is Vowel or Consonant
If YES GOTO STEP 8 IF NO GOTO STEP 3
8: CHECK FOR END of string
If No GOTO STEP 3
9: Compute Character
10: STOP

5.4 Digital Speech Pronunciation

The recording was done by a male as the native speaker of the language. Integrating the recorded syllable sound to match the block of supplied text so as to have correct pronunciation was achieved with JAVA programming language and implemented the design of a text to speech plug-in for Mozilla fire fox browser.

5.4.1 Speech Pronunciation Algorithms

STEP

1: START
2: SET COUNTER=1
3: DECLARE VALUE for n
4: Check if Char[count]
5: If No GOTO STEP 8
6: BREAK Char is space
7:GOTO STEP 10
8: Check if the Char is space
If No GOTO STEP 10
9: BREAK SPACE //{Pronounce the word}
10: SET COUNTER= COUNTER+1
11: STOP

6. SYSTEM IMPLEMENTATION AND EVALUATION

6.1 Program Development

Program development is simply concerned with how to provide an efficient (economic) and effective (relevant and useful) system.

It is an integral part of software development embarked after a detail analysis of the system has been done and the project feasibility study undertaken. The purpose of the design is to meet the user's specification of the system's software and determine flexible system alternatives that will achieve the recommended result and make optimum use of the hardware, software and other processing resources that may be useful in the implementation and solution finding.

For there to be good system design, top down development was used so as to achieve an order that can lead to decomposing the requirement into well specified high level model.

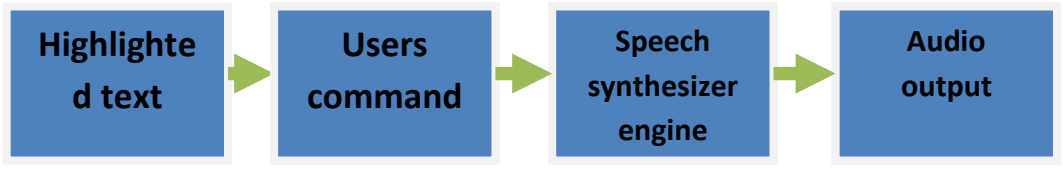
6.1.1 Program Development Tools

DOI: 10.31695/IJASRE.2019.33568

The language used in designing this program is node.JS in conjunction with some plug-in development API(s). This language was chosen to design a text to speech extension because it offers a great deal of plug-in development tool and it is very easy to understand.

6.2 Text To Speech Conversion Modules

This module involves the highlighting of text on the browser. This highlighted text is then parsed at the click of the speak option that pop up when the user right click on the highlighted text. The speech synthesizer engines grab the highlighted text and perform a series of actions on it after which it will return the audio output of the text.



Other functions added to this application are play, pause, stop and download.

- Play: this function allow the user to resume the play of the audio file when paused.
- Pause: this function is use to pause the audio output.
- Stop: this is used to terminate the audio output and also to exit the text to speech extension interface.
- Download: this function enable the user to download the audio output so that it can be reuse later.

6.3 System Requirement

This system requires hardware, software, human-ware which enables the system to work.

6.3.1 Software Requirements

Software is a program used by computers to facilitate their operation and utilization. It gives the computer the capability of doing whatever the user has the need for. A computer without software is like an empty box.

6.3.2 Hardware Requirement

Hardware is the physical equipment or components that make up the computer system. It refers to the physical interface of the component that can be seen and touched. Every software has minimum hardware requirement needed for its operation. The minimum hardware requirements required for the text to speech application are;

1. Central Processing Unit(CPU)
2. Visual display unit(VDU)
3. Random Access Memory (RAM) at least 64MB memory space.
4. Hard-disk (At least 2.4GB of storage space)
5. Keyboard and mouse
6. Audio hardware (headset)

6.4 System Interface

Every program has input interface as well as output interface. They are used mainly to achieve specific objective of verifying the processing operation been performed.

6.4.1 Input Interface

The input interface describes the manner in which data enter the system for processing. The input interface for the proposed system is a part of the program where the user highlights the text i.e. the input to be converted to speech with the aid of the mouse.

DOI: 10.31695/IJASRE.2019.33568

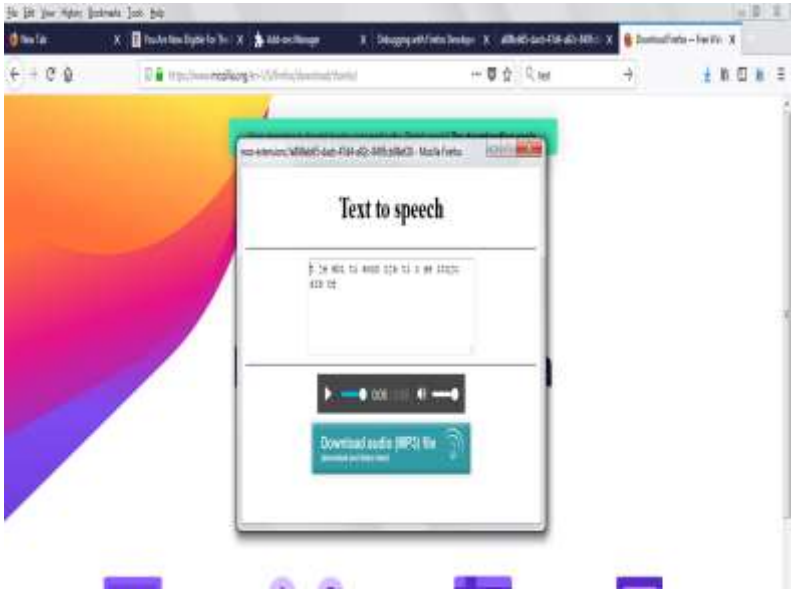


Fig 6:1: Output Interface Of The Program

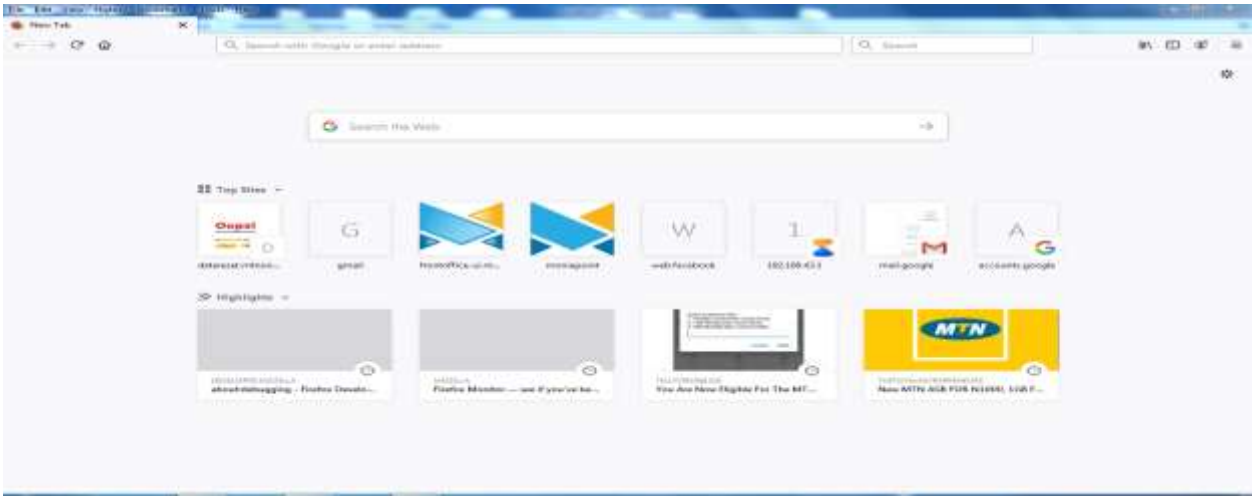


Fig 6.2: Lunching of the Mozilla Firefox to upload the text to speech source code

DOI: 10.31695/IJASRE.2019.33568

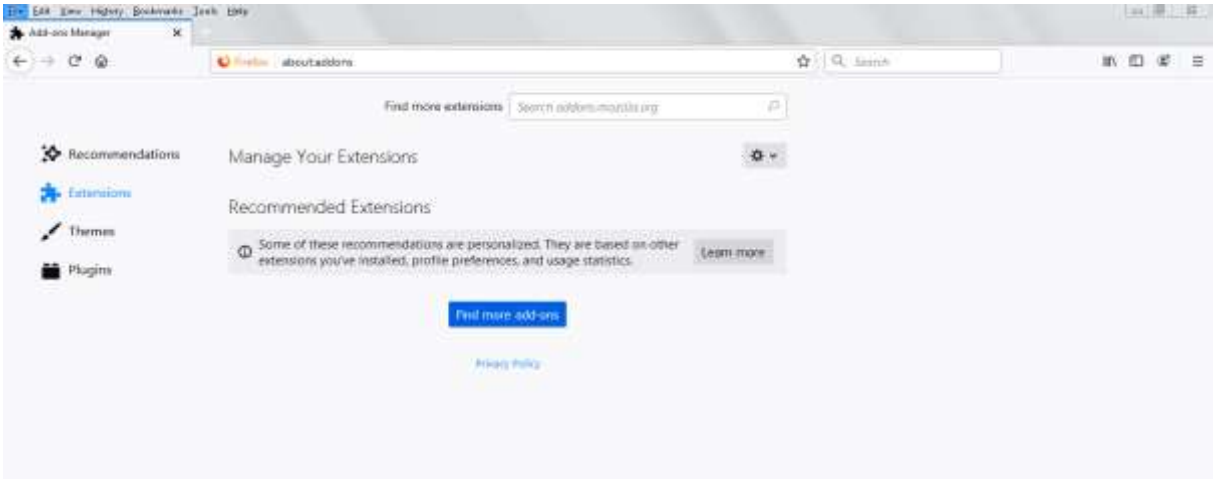


Fig 6.3: Interface of file extension to upload the java extension of the text to speech

Click on the setting icon by the left top corner of the page

Drop down menu display

Click on Debug Add-ons

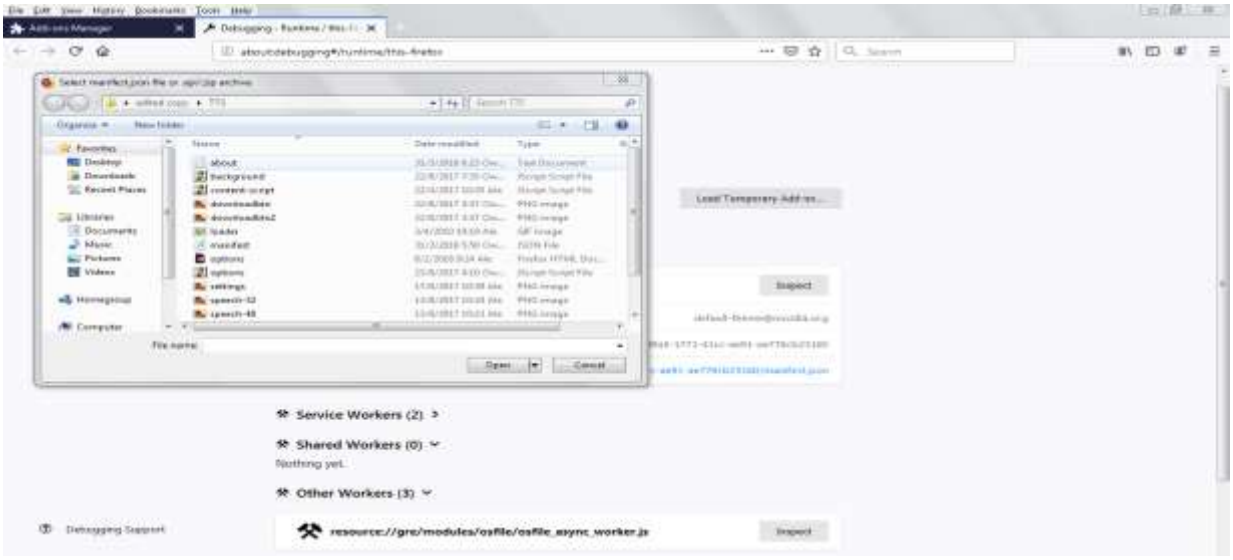


Fig 6.4: Click on load temporary Add-on to load manifest code of the program text to speech

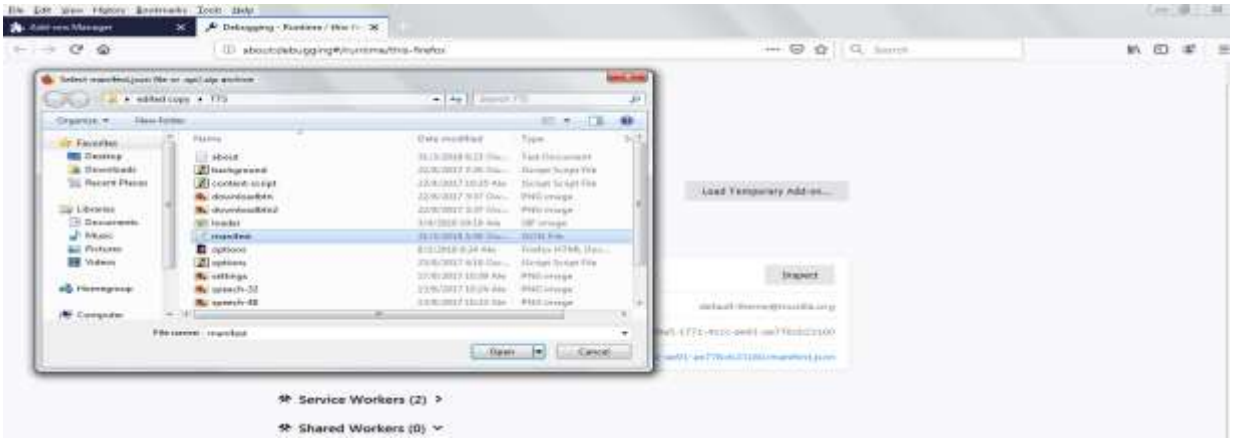


Fig 6.5: Select manifest and click on open

DOI: 10.31695/IJASRE.2019.33568

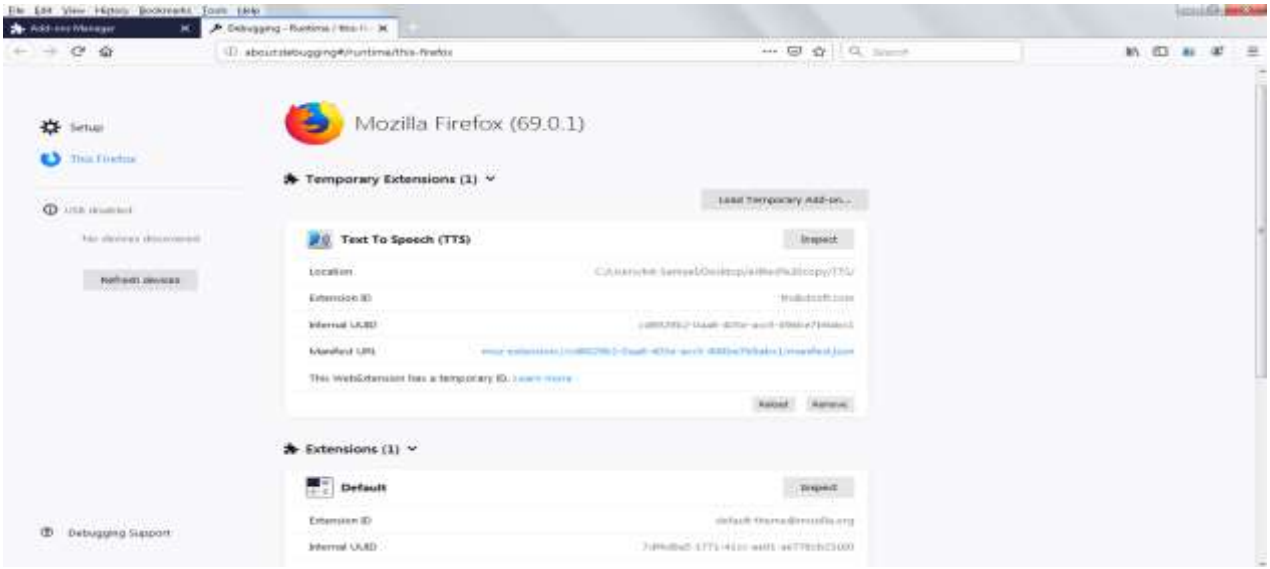


Fig 6.6: Uploaded successfully

7. SUMMARY, RECOMMEMDATION AND CONCLUSION

7.1 Summary

The developed system gives a very simple method for the text to speech conversion for Mozilla Firefox browser. Text to speech conversion is achieved and received a better result which is audible and perfect. A text to speech extension is an application that converts the highlighted text into spoken word by analyzing and processing the text. The user of the text to speech is to read out the highlighted text to the user which can be saved as an audio file. The developed text to speech extension will be of great help to people with lack of pronunciation skill.

8. RECOMMENDATION

It is recommended that this plug-in to be used by all internet cafés. The plug-in should be installed in all Mozilla browsers. They will find this plug-in useful to aid their service delivery to customer.

It is also recommended that this plug-in should be installed in all homes and offices with Mozilla browser. This plug-in can also improve the learning skills for the young Yoruba language learners and also aid learners lacking or deficient in pronunciation skills.

9. CONCLUSION

Text to speech synthesis is a rapidly growing aspect of computer technology and it is increasingly playing more important roles in the way we interact with the system and interface across a variety of platforms. In the course of this study, the researcher has identified the various operations and processes involved in text to speech synthesis. The researcher has also developed a very simple and attractive graphical interface. The system interfaces with the text to speech engine developed.

REFERENCES

[1] Bhusan, B. Krishna - Nepali Text to Speech Synthesis System using ESNOLA Method of Concatenation, International Journal of Computer Application, vol. 62: 24-28, 2013.

[2] Sproat R, Olive (1999). “Multilingual Text-to-speech synthesis: The Bell Labs Approach.

[3] Sasirekha, E. Chandra - Text-tospeech: a simple tutorial. International Journal of Soft Computing and Engineering, vol 2, 275-279, 2012.

[4] Sher Y.J, Chiu, M.-C. Hsu, K, C. Chung - Development of Hmm Based Taiwanese Text-To-Speech system. In International Conference on Software Techniques and Engineering, 330-333, 2010.

DOI: 10.31695/IJASRE.2019.33568

[5] Zeki M, O. Othman, A. Khalifa, W. Naji - Development of an Arabic Text-To-Speech System. In International Conference on Computer and Communication Engineering (ICCCE), 11-14, 2010.

[6] Sak H, Gungor T, Safkar Y - A Corpus Based concatenative speech synthesis system for Turkish. Turkish Journal of Electrical Engineering and Computer Science, vol. 14: 209 – 223, 2006.

[7] Sadaoki, F. K. Tomonori, S. Yousuke, H. Chiori - Speech-to-Speech and Speech-to-Text Summarization. First International Workshop on Language Understanding and Agents for Real World Interaction. 2003.

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

[9] Alan, W. (2002): "Perfect synthesis for all the people all of the time". Keynote address at the Workshop on Text to Speech on 30th of Seat: http://www.cs.cmu.edu/

[10] 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

[11] Afolabi A.,(2012): "Development of Yoruba Text To Speech System" Ladoke Akintola University of Technology, Ogbomoso, Nigeria.

[12] Boroş, T. (2013). A unified lexical processing framework based on the Margin Infused Relaxed Algorithm.

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

[14] Mogey N., (1999): "So you want to use a Likert Scale?" Learning Technology Dissemination initiative, HeriotWatt University.

DOI: 10.31695/IJASRE.2019.33568