International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12– No. 39, April 2022 – www.ijais.org

A Grapheme-based Text to Speech System for Yoruba Language

Itunuoluwa Isewon
Department of Computer and Information Sciences Covenant University PMB 1023, Ota, Nigeria

Adejumoke Famade
Department of Computer and Information Sciences Covenant University PMB 1023, Ota, Nigeria

Jelili Oyelade
Department of Computer and Information Sciences Covenant University PMB 1023, Ota, Nigeria

ABSTRACT
A Text to Speech synthesizer is an application that converts text to speech; the process of conversion is achieved by the application of the analysis of natural and digital signal processing on the input text. Over the years, some Text to Speech systemshas been built for different languages e.g. English, Kiswahili, German, French, Telugu, Mandarin, etc. The aim of this study is to develop an application for Yoruba Text to speech synthesis.

The Mary Text to Speech framework was used to implement a speech synthesizer for the Yoruba language. The Text to Speech system developed can convert an input text into speech sound in Yoruba. This application allows a user to input a text and the engine reads out the text as a synthesized speech. The user is also able to upload a text file and the engine also reads the uploaded text file, the user can save the synthesized speech inthe local storage ofthe system.

A basic graphical user interface has been produced that plays out the essential elements of a speech synthesizer like conversion and speech synthesis. Agrapheme-based speech synthesizer has been produced for the Yoruba language which is a tonal language.

Keywords
Text to Speech, Speech synthesis, Mary tool, Yoruba language

1. INTRODUCTION
The process of automatically converting a text passage into speech sound like the native speaker is called Text-to-Speech (TTS) synthesis. Computers have the capability of communicating with users via aTTS synthesizer. The text is entered into the TTS system and a computer algorithm that is known as the TTS engine analyses it, pre-processes it, and the speech is synthesized with some mathematical models. The TTS system usually produces sound in an audio format as the output [1]. The TTS synthesis technique consists of two main phases. The first is text analysis, where the input text is transcribed into a phonetic or some other linguistic depiction.The second one is the production of speech waveforms, where phonetic and prosodic information generates the output. These two phases are regularly called high and low-level synthesis [2]. The various examples of the input text include data from a word processor, standard ASCII from an e-mail, scanned text from a newspaper, or a text message from a mobile device. A pre-processed character string is analyzed into a phonetic representation that is usually a string of phonemes with some additional information for the authentic intonation, duration, and stress. The low-level synthesizer generates speech sounds with the information from the high-level synthesizer. There is a history of the development of synthetic speech-like sounds, with documented mechanical attempts dating to the eighteenth century.

There are two categories of languages: tonal and non-tonal language [3]. Yoruba, Mandarin, Hausa and Igbo are few examples of tonal languages while Spanish, Korean, German, Polish, and French are examples of non-tonal languages. As at 2021, there are about 38 million native speakers of the Yoruba language in three African countries (Nigeria, Benin, and Togo). Not much has been done with respect to the development of a TTS engine for Yoruba. Hence, there is a need to implement a TTS framework for the Yoruba language [4]. A Yoruba TTS system is important for the Yoruba-speaking people who are physically impaired and have little or no knowledge ofthe English language. This system is also a platform for people who want to learn how to speak Yoruba in other words it can be a learning platform for the non-Yoruba speaking people.

2. SPEECH SYNTHESIS
2.1 Overview
The generation of speech is produced by regulating the airflow from the lungs to the throat, nose, and mouth. Sound is said to convert the symbolic representation of what to say into an actual speech waveform [5].An objective of a TTS system is to convert raw textinto speech form. Speech synthesis is the synthetic production of human sound where text is automatically changed into phonetic data and then converted into an acoustic waveform [6].

The present-dayTTS system changes over text into a 'produced speech'sound in these two main phasesi.e., High-Level Synthesis (HLS) reads the speech sounds and creates an illustration of the desired acoustic signal. The HLS phase is actualized utilizing two components, the principal component i.e.,the Text examination component which investigates the stored text to distinguish its fundamental components and the setting whereby they are utilized. However, the results of the Text examination component are passed into the next component which is; the prosody component, which helps to maintain expressiveness and intelligibility in speech synthesis. Prosodic components are Pitch, duration, accent, and phrasing. All preparation required in this phase is termed High-level synthesis (HLS) and the innovation for actualizing the speech is measured from the space of Natural Language Processing (NLP) [7].

If the contribution to a speech synthesizer is specified as text, the system is known as a TTSsynthesizer. Nonetheless, because of speech synthesizers with confined vocabulary, for example, technologies playing pre-recorded speech samples, the description is clear. TTS synthesizer contains two parts, called the abnormal state and low-level combination. The abnormal state synthesis changes over the text commitment to an edge that thinks about the needed acoustic phonation of the verbalization. This suggests changing over the text commitment to a phonetic or some other etymological illustration and foreseeing the prosody. At the same time, the data text is at first institutionalized into natural words, and the source characteristics of the text are examined. Therefore, the text is changed over to a phonetic level, this is called TTS change [8].

2.2 Speech production method

The most critical characteristics of a speech synthesis system are naturalness and the ability for the users to understand. Naturalness depicts how intently the speech resembles the human sound, whereas clarity is the straightforwardness wherefore the speech is understood. Speech synthesis systems as a rule attempt to amplify both attributes [9]

Format Synthesis: Format synthesis is driven by principles; it produces speech signals from sound parameters that are inferred by human speech specialists from speech information. Format speech synthesis does not make utilization of human speech tests at runtime. Rather the synthesized speech sound is made by utilizing added substance synthesis and sound model (physical demonstrating synthesis) [10]. This is the most vital acoustic speech synthesis technique.Format synthesis utilizes the primary-channel hypothesis of speech generation. Diverse phonemes are built by changing the inside recurrence, data transmission, and pick up of every channel. The basis can be demonstrated with speech heartbeats or commotion. Today, the nature of format synthesizers is sub-par in contrast with the most recent synthesis techniques, for example, concatenative and LPC-centered strategies.However, format synthesis has numerous applications in perusing technologies for the visually impaired and in speech observation tests for making bots [10].

Vowels are synthesized by passing an occasional sound flag through a channel in the voiced sounds. For unvoiced speech sound, it is generally displayed as background noise, format based includes both serial and parallel format synthesizers. The image grouping is changed over into an arrangement of parameter vectors in the format-based synthesizer. Parallel formats permit singular sufficiency to be controlled that is the reason it is likened to serial formats.Format synthesis is efficient with low implementation cost, the naturalness of the output speech sound is limited, and it is quite difficult to extract the parameters for speech data.

Articulatory Production: Articulatory production attempts to model the normal speech production preparationIt is hypothetically the finest technique for brilliant speech synthesis; nevertheless, it is the most complex way to deal with speech synthesis and its usage. Due to the confinements of the present speech creation representations and computational control, compared to other speech union procedures, articulatory synthesis has not gained as required ground. Regardless, it has various supportive applications in essential speech study [11].

LPC-Based Production: The source filter model of speech production is used in linear predictive coding (LPC) the same way as in format synthesis, instead of finding the parameters for individual formant filters the filter coefficients can be automatically estimated from a short frame of speech in the LPC-based synthesis. The filter coefficients is used to synthesize speech with an appropriate excitation. Dependent on whether the synthesized speech segment is voiced or unvoiced, excitation can be either a periodic source signal or noise. Linear prediction (LP) is a widely used technique in speech technology. Linear prediction (LP) is a widely used method in speech technology. However, the quality of a basic LPC vocoder is considered poor, high quality synthetic speech can be produced with more sophisticated LPC-based synthesis methods [11].

Sinusoidal Synthesis: Sinusoidal synthesis decomposes each frame in a harmonic model into a set of harmonics of an estimated frequency fundamental frequency. It changes the fundamental parameters like amplitude, frequency, and phases by keeping the same spectral envelope. Modeling every spectral component as a sinusoid is the basic idea behind a sinusoidal synthesis [6].

Concatenative Synthesis: Recorded samples of real speech are smoothly joined to generate an arbitrary synthetic sound in the Concatenative synthesis. Common unit lengths are words, syllables, demi syllables, phoneme, diaphone, and triphone. Concatenative synthesis can generate highly comprehensible and natural synthetic speech because the natural features of the speech are preserved in the units. The set of speech units is always limited; the discontinuities in concatenation points can cause distortion despite the use of various smoothing algorithms. Concatenative speech synthesis is less flexible because. It is very impractical or impossible to store all the necessary units for various speakers in various contexts. The need for vast storage for all the recorded units is another disadvantage, but with the cost of computer storage decreasing, and with the development of fast database access techniques, this problem is not as stern as it used to be. Concatenative speech synthesis is generally used but might not be the best solution because of the mentioned limitations [11].

Hidden Markov Model Synthesis: Hidden Markov Model is one of the widely used methods in speech synthesis. Hidden Markov Model (HMM) is a factual model, which can be utilized for demonstrating the speech parameters extricated from a speech database, and after that producing the parameters as indicated by text contribution for making the speech waveform. Speech is delivered in different talking styles with various speaker qualities and even feelings utilizing the Hidden Markov Model speech synthesis. HMM additionally benefits from better flexibility and unmistakably littler memory necessity. Nevertheless, the HMM-based TTS frameworks regularly experience the ill effects of corrupted instinctive nature in quality contrasted with concatenative-based speech synthesizers. Regardless, the HMM-based Text-to-Speech frameworks are growing quickly, and much work is completed for finding systems to improve the quality and expectation of engineered speech. The current pervasive stage for HMM-based speech combination is the HTS framework created in Japan (HTS 2008) [11].

2.3 Yorubaphonology

The Yoruba letters comprise 25 letters that were gotten from Latin characters [6], of which 18 are consonants and the rest

are vowels. This is illustrated in tables 1 and 2 below. Yoruba and English vowels share similar features due to their origin; the features include Vowel sounds are usually voiced compared to consonant sounds, the articulation of vowel sounds varies because of the opening of the mouth and tallness of the tongue, and vowel sounds produced require an unrestricted flow of air through the mouth.

Table 1: The capitalized and lower case representation of Yoruba letters in order [12]

Aa	Bb	Dd	Ee	Èè
Ff	Gg	GBgb	Hh	Ii
Jj	Kk	Ll	Mm	Nn
Oo	Òò	Pp	Rr	Ss
Šš	Tt	Uu	Ww	Yy

Table 2: Orthography Illustration of Yoruba consonant [12]

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

3. RELATED STUDIES

Various Text to Speech systems has been created. With sorts of systems utilized. Likewise, various other languageshas been created with various synthesizer strategies that looked into their framework and their basic innovations [12].Gakuru and colleagues' [13]work presented here is the change of a Kiswahili Text to Speech System (TTS) in the perspective of the Festival Unit Selection Speech Synthesizer. The system made here is concatenative, which suggests that amalgamation is done through making waveforms by interfacing parts of standard exchange recorded from individuals. German Text-to-Speech framework developed by Charfuelan*et al.* [14] utilizes the MARY (Modular Architecture for Research on speech synthesis) which is an adaptable instrument for research, improvement, and education in the space of TTS synthesis [14]. Though Mary TTS was originally developed for the German language; nowadays it makes available voices and support for the following languages: US English, British English, German, Turkish, Russian, and Telugu [14]. Concatenative speech synthesis was utilized in Hindi Text-to-Speech implementation. The method involved three basic steps:TextPre-processing, Text Processing, and Speech synthesis [15].

4. SYSTEM ANALYSIS,DESIGN AND IMPLEMENTATION

The Yoruba language TTS system was developed using Java studio, MARY TTS framework, and NetBeans. The Mary (Modular Architecture for Research on speech synthesis) Text-to-Speech synthesis system is a flexible and modular tool for research and development in the domain of Text-To-Speech synthesis [16]. Mary TTS is an open-source project, it is written in Java and includes a few useful tools for adding support for a new language. Mary supports the production of new voices starting with no outside help, that is, it gives the important instruments and non-exclusive reusable run-time framework modules for including a language. Other technologies and tools used in the implementation of the system are: .Net framework, Microsoft Speech, Natural Language Processing, and Maven. A grapheme-based desktop application was developed to be able to convert an input text into a synthesized speech form. The prompts were recorded and synthesized to be able to produce graphemes of the language. Speech produced using Mary appeared to be preferable to the speech sound producedon the windows applications. The figures 1 - 3 show the systematic procedure of how the TTS system was implemented. The sequence diagram (Figure 2) depicts the flow of messages and the interaction of objects in the TTS system. A user inputs a text on the machine interface, the interface then communicates with the speech engine that there has just been a text input on the interface, the engine then waits for the next instruction, when the user clicks Play, there is another communication between the interface and the speech engine. The speech engine then carries out the instruction and speaks what the interface has communicated.

The Yoruba Text-To-Speech systemuser can convert text to speech either by typing the text into the text field provided, by coping from an external document in the local machine and then pasting it in the text field provided in the application or by uploading a file from the computer or local storage device.The precondition is that the user must type in a valid text document for the interface to convert the text to a synthesized speech (Figure 2). The post-conditions are either a success end where an audible speech is produced or a failure end where the system display an error message and prompts the user to input a new text. The Yoruba TTS system has an exceptional functionality that gives the user the option of saving its already converted text to any part of the local storage drive or the desktop in an audio format; this allows the user to copy the audio format to any of his/her audio devices.

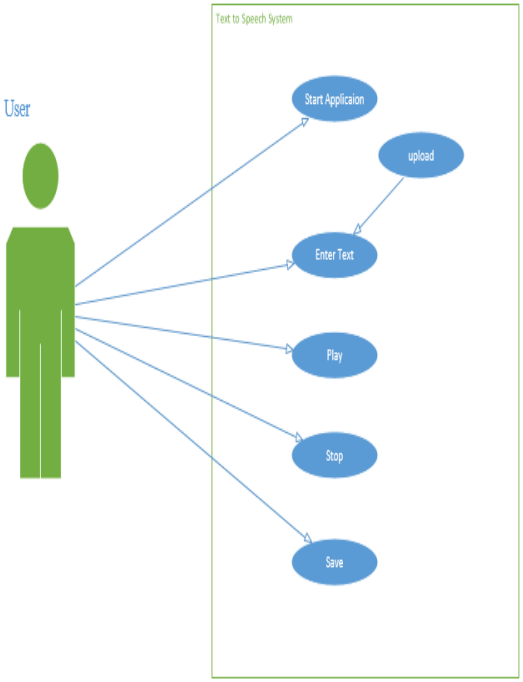

Figure1: A use case diagram of a Text-to-Speech system

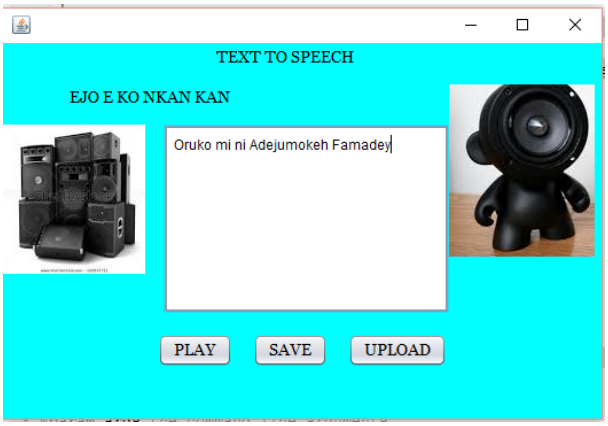

Figure 3: User interface for the desktop application

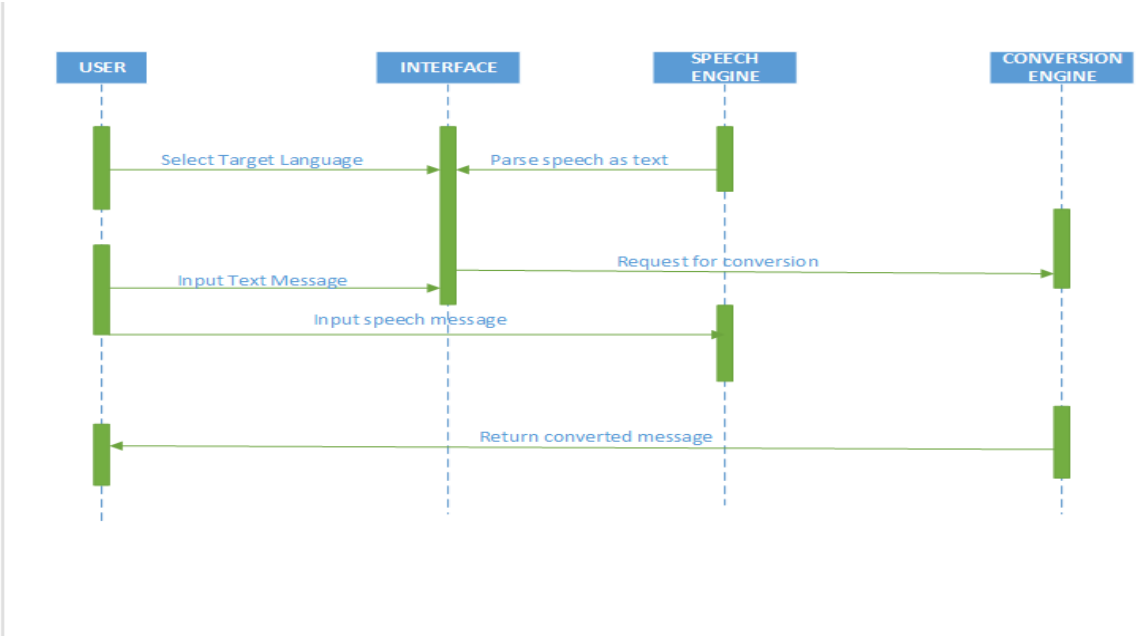

Figure2: Sequence Diagram of the Text to Speech system

Main User Interface: this platform is where all the activities are carried out. The user can choose to upload a text instead of manually inputting it; the user can choose to save the synthesized speech. Finally, the user can write in the text area and press "PLAY" to listen to the speech.

Save Interface: With this platform, a user can save the synthesized speech (Figure 4).

Upload Interface: A user can upload a text in this platform instead of manually typing in the texts (Figure 5).

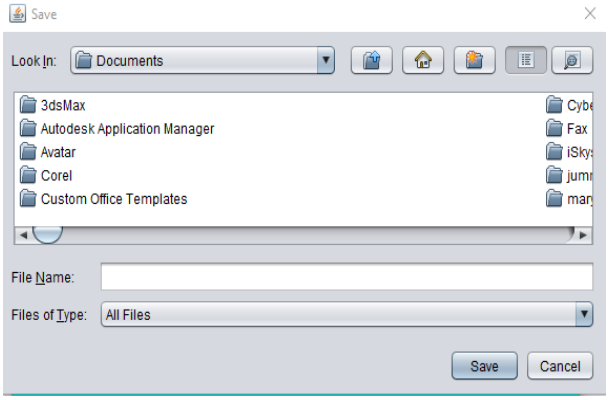

Figure 4: JAVA dialogue box to save the synthesized speech

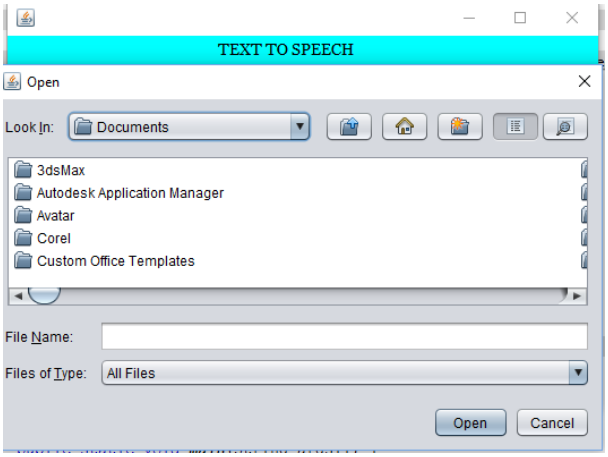

Figure 5: JAVA dialogue box to upload a text file

5. CONCLUSION

In this study, the diverse sorts of speech synthesis techniques have been discussed, and the procedures and steps required in speech synthesis havealso been highlighted. A basic graphical user interface has been developed that plays out the essential elements of a speech synthesizer like conversion and speech synthesis. Likewise, a grapheme-based speech synthesizer has been implemented for the Yoruba language. In the future, machine learning algorithms should be used for the development of a speech synthesizer from scratch to design the basic features of speech sound for a wide range of other Nigerian languages. Another area of further work is the implementation of a text-to-speech system on a mobile-driven system and a web-based system. Additionally, the TTS system should be developed in such a way that it will be able to also convert Speech to Text and translate the inputted text from a specified language to another language for better learning and understanding.

6. REFERENCES

[1] Isewon, I., Oyelade, J., &Oladipupo, O. (2014). Design and Implementation of Text to Speech Conversion for Visually Impaired People. International Journal of Applied Information Systems, 7(2), 25–30.

[2] Dutoit, T., 1993. High quality text-to-speech synthesis ofthe French language. Doctoral dissertation, FacultePolytechnique de Mons.

[3] Duanmu, S. (2004). Tone and Non-Tone Languages: An Alternative to Language Typology and Parameters. Language and Linguistics, 5(4), 891–924.

[4] Ekekwe, N. (2015). Development of Text to Speech Applications. Covenant University.

[5] Byrne, B. (2014). Engineering Part IIB : Module 4F11 Speech and Language Processing Lectures 9 & 10 : Weighted Finite State Transducers for Speech and Language Processing.

[6] Oloruntoyin, S. T. (2014). DEVELOPMENT OF YORUBA LANGUAGE TEXT-TO-SPEECH E-LEARNING SYSTEM. International Journal of Scholarly Research Gate, 2(1), 19–36.

[7] Shih, C., &Sproat, R. (1996). Issues in text-to-speech conversion for Mandarin. Computational Linguistics and Chinese Language Processing, 1(1), 37–86.

[8] Polomé, E. C. (1967). Swahili language handbook.

[9] Laver, J. D. M. (1970). The production of speech. New Horizons in Linguistics, 53–75. https://doi.org/10.1007/978-1-4613-8202-7_6

[10] Sarroff, A. M., & Casey, M. (2014). Musical Audio Synthesis Using Autoencoding Neural Nets. Proceedings of the International Computer Music Conference, 1(September), 14–20.

[11] Raitio, T. (2008). Hidden Markov Model Based Finnish Text-to-Speech System Utilizing Glottal Inverse Filtering. Master's Thesis, HELSINKI UNIVERSITY OF TECHNOLOGY. Retrieved from http://lib.tkk.fi/Dipl/2008/urn012274.pdf

[12] Afolabi, A., Omidiora, E., &Arulogun, T. (2013). Development of text to speech system for yoruba language. In Innovative Systems Design and Engineering—Special Issue of 2nd International Conference on Engineering and Technology Research (Vol. 4, pp. 1–8).

[13] Gakuru, M., Iraki, F. K., Tucker, R., Shalonova, K., & Ngugi, K. (2005). Development of a Kiswahili text to speech system. Interspeech-2005, 1481–1484.

[14] Charfuelan, M., Pammi, S., Steiner, I., Pierre, U., Upmc, C., & Tts, M. (2013). MARY TTS unit selection and HMM-based voices for the Blizzard Challenge 2013.

[15] Kamble, K., &Kagalkar, R. (2014). A Review : Translation of Text to Speech Conversion for Hindi Language, 3(11), 1027–1031.

[16] Schröder, M., &Trouvain, J. (2003a). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. International Journal of Speech Technology, 6(4), 365–377.

[17] Schröder, M., &Trouvain, J. (2003b). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. International Journal of Speech Technology, 6(4), 365–377. https://doi.org/10.1023/A:1025708916924

[18] Schröder, M., &Trouvain, J. (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. International Journal of Speech Technology, 6, 365–377. https://doi.org/10.1023/A:1025708916924.

[19] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., &Alku, P. (2011). HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering. IEEE Transactions on Audio, Speech, and Language Processing, 19(1), 153–165. https://doi.org/10.1109/TASL.2010.2045239