Graphic User Interface for Hausa Text-to-Speech System

Umar Adam Ibrahim
Computer Science
Nile University of Nigeria
Abuja, Nigeria
umaradamibrahim@nileuniversity.edu.ng

Moussa Mahamat Boukar
Computer Science
Nile University of Nigeria
Abuja, Nigeria
musa.mohammed@nileuniversity.edu.ng

Muhammed Aliyu Suleiman
Software Engineering
Nile University of Nigeria
Abuja, Nigeria
muhammad.suleiman@nileuniversity.edu.ng

2022 2nd International Conference on Computing and Machine Intelligence (ICMI) | 978-1-6654-7483-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICMI55296.2022.9873676

***Abstract*—Natural language processing and Digital signal processing are broadly used methods used to enable systems to understand commands and manipulate speech or text. Most of the Text-to-speech done was for major languages such as English, French and others, with no or little for African languages like Hausa, which are termed under resource languages. In this paper, we developed a graphical user interface for the Hausa Text-to-Speech system. This system converts Hausa text to Hausa audio sound, by processing and analyzing it using natural language processing and Digital Signal Processing. Our graphical user interface, aid in converting entered Hausa language text into Hausa speech.**

Keywords—Hausa, Text-to-Speech, Interface, System

I. INTRODUCTION

In human Intelligent System, the primary means of communication happens to be via speech. Natural language processing has served as a major component in several field of language computing. Researchers have carried out so many work on the application of NLP. The researches include how can computer understand, identify and manipulate natural language. Another application is converting Textual data to Speech data; this is called text to speech application. Transforming from text data to speech data is the automatic conversion of text to speech. Hence, TTS is the technology that enable user to speak to computer by entering text and allows computer to speak back to user. TTS system gets the inputted text, then a TTS algorithm analyzes the text, the text is then pre-processed. Then speech synthesized is now applied on the processed text based on some statistical methods. Finally, the TTS machine produces the output sound in an audio format.

Presently, there are well-developed TTS system available for major languages and some for under resource languages. Such as [1]. There are even advance systems such as STS [2] which is a system for virtual impaired people. As [3] introduced optical character recognition process in TTS system. Furthermore, other work includes text in image as in [4], which can be track by finger camera. However, for local languages as presented in [4] applied concatenation method.

It is crucial to understand that, TTS has been developed to support people with challenge such as blind people, people with reading challenges, dyslexia people and others [5]. One interesting aspect of speech system is that it can either be a software or hardware device. Very significant amounts of work have been done to improve this field. This has led to the creation of high quality Text To Speech systems for major languages which some are commercially available for use.

The process of changing written word to spoken word is nontrivial. However, the sound intonation generated needs to be natural and original. Speech synthesis cannot be achieved by cutting and pasting smaller units together. It required analyzing and processing. The processing stage in speech signal is of paramount important, there is need to focus more on smoothing to achieve natural speech. And according to [6] speech system is classified into three categorize: which are Articulatory synthesis, Format synthesis and Concatenative synthesis.

There are few NLP research work for under-resourced languages such as Hausa and other Africa languages. This is due to language barrier and lack of text dataset from such languages. With this challenges and barriers, our effort is to build a Graphic User Interface for Hausa Text-to-Speech system. This GUI would allow user to enter or upload Hausa text. The system would convert the entered text or file into spoken Hausa word.

Lastly, the work started with the introduction in section I, section II related work on Text-to-Speech systems, section III presented Text-to-Speech synthesis model. Section IV outlined the implementation method and section V is the conclusion.

II. FROM TEXT DATA TO SPEECH DATA

Presently the deployment of Text to Speech machine is advancing at a very rapid rate. To create a Text To Speech platform the following needs to be put in place: word segmentation, concatenation and extraction.

Kannada language is a language spoken in south India [5] developed a text to speech platform for Kannada language. The collected text was recorded, thus the recorded speech was further segmented into unit words. These unit words were stored in a database. Using MATLAB 2010 platform the stored units' word were concatenated to build a TTS platform for Kannada language.

[6] designed & developed a Text to Speech platform for Bengali language. The system major features are wave synthesis, prosodic analysis, and Phonetic analysis and text normalization. For experimental purpose, external people were engaged to write down what they hear from the system. The result shows that the accuracy of the word level is 73.3% and 93.3% for sentence level rate.

Authorized licensed use limited to: Wikipedia. Downloaded on January 20,2025 at 09:39:54 UTC from IEEE Xplore. Restrictions apply.

In [7] a text to speech platform for Arabic language was implemented. The authors built their system by implementing a method that concatenant allophone and diphone. The system comprises of two modules: text and linguistic analyzer and synthesizer.

In [8] Taiwanese speech system was developed by creating a sub-syllables units database. The database composed of multiple accented corpuses. The corpus is made up of basic Taiwan phonemes, vowels and consonants.

[9] developed a rule-based algorithm to transform written text to spoken word. A hybrid system combined formant and concatenation methodology. The experiment shows a promising output of 96% in recognizing word, phrase and sentences.

The Indian English language synthesis was developed based on HMM algorithm [10]. The system was built with 1000 phonetically studio recorded Indian English sentence.

Improving synthesized output for text to speech synthesis is a crucial stage, however Viterbi coder methodology had great impact in optimizing speech synthesizer [11]. Viterbi coder algorithm uses memory access techniques along with pipelined precomputation. The precomputation reduces the power consumption of the memory.

Of recent End-to-End algorithm has been used in developing speech technologies [12]. This research implemented an end-to-end text-to-speech application.

Built text-to-speech system from raw data set is not an easy task. [13] proposed an automatic data audio data processing method for Bangla text to speech system. In [14] the developed a voice conversion framework by implementing a sequence to sequence multi-speaker text to speech synthesis. The developed a platform that outperformed the tradition methodologies, which comprises of phonetic posterior gram and Auto VC methodologies.

The quality of TTS system has always been a challenge. However, [15] worked on improving the quality of TTS. The authors implemented neural network based on supervision of perceptual loss. The advantage of the proposed method is that it is independent regardless of the model architecture. The achieved mean opinion score and the phone error rate indicated some improvement achieved by the proposed system compared to the previous method. In [16] the author implemented text-to-speech by applying incremental text-to-speech method. The goal is to achieve speech quality and also decrease latency. The incremental TTS was implemented with the aids of pseudo look ahead, which was generated with a large language model.

Of recent, most of the implemented TTS are based on neural methodology. Neural methodology proved to be more superior compared to conventional statistical methods [17]. However, the challenge of performance degradation still occurs in situation where we have out-of-domain test dataset. Furthermore, the challenge of bias problem is another challenge. Thus [17] solved the above stated challenges by building multi-teacher knowledge distillation (MT-KD). MT-KD solved the out-of-domain and bias problem. It's outperformed the data augmentation and adversarial learning methodologies respective.

In [18] the author presented normalization processing module for the development of Luganda system for the conversion of text to speech. The normalization was done by implement a rule-based method for the verbalization, detection and classification of Luganda text. After running the system with 7 dataset the average detection rate was 82% and the normalization rate was 77.7%.

III. METHODOLOGY

Generally, the development of a Text to Speech system passed through several phases. The features include selection of text unit, text unit normalization, preprocessing, text-to-phone and grapheme-to-phoneme. Furthermore, Speech system is implemented in several ways like Format, Articulatory, HMM based, Sinewave Synthesis and Concatenative synthesis. Our Hausa-TTS is developed based on following procedures:

A. NLP Module:

Define as the natural language processing component that generates the phonetic transcription of the read text. It's helps in generating speech synthesis method as shown in figure 1.

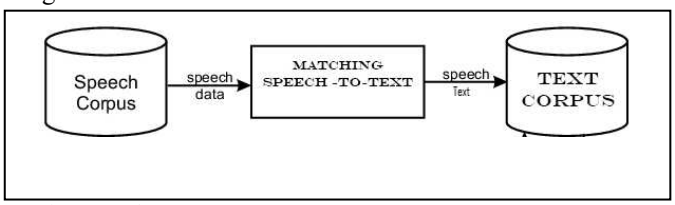

Fig. 1. Database

B. DSP Module

This is a Digital Signal Processing module, this propagate the symbolic representation that was received from NLP into audible sound known as Natural Language Generation NLG as shown in figure 2.

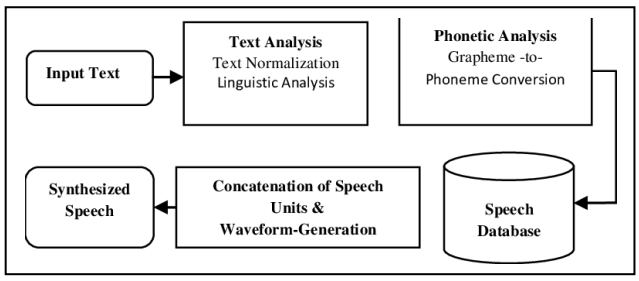

Fig. 2. Text-to-Speech Model

The software was designed and developed using a simple application. The application was built using python 3.8 version. The platform is segmented into two major modules:

C. Interface Module:

This module contains the basic home interface which manages the primary operation of the application such as typing or pasting text directly.

Authorized licensed use limited to: Wikipedia. Downloaded on January 20,2025 at 09:39:54 UTC from IEEE Xplore. Restrictions apply.

D. *TTS Engine*

This module analyze the entered text, matches the entered text with the available recording in the database for conversion.

The application converts the entered text to speech. Input text can be entered via

- Keyboard: that is by typing directly from the keyboard.
- Text Pasting: That is by copying and pasting the text into the open box in the application.

IV. EXPERIMENT

This comprises of text preparation, implementation and performance validation.

A. *Text Corpus Preparation*

This comprises of collection, cleaning and labeling of the text data. The Text corpus were acquire from Hausa online News outlet. These are public available newspaper. The collected text covers daily human activities. The features of the corpus is outlined in table I.

TABLE I PROPERTIES OF THE TEXT

Corpus	Number	Distribution	Percentile
words	100,000	Affirmative sentences	80%
Sentences	2,000	Interrogative sentences	35%
		Exclamation sentences	10%
		Avg word/sentence	11.3%

B. *Corpus Evalution*

To check the quality of the corpus. An evaluation of corpus was done based on “Unit selection algorithm”. The authors implemented the mean opinion score (MOS) for evaluation purposes. The general output of the text corpus is result is presented in table II.

TABLE II. RESULT OF MOS EVALUATION

Grade	Point	Average score (%)
Unknown	0	15
Poor	1	12
Fair	2	20
Good	3	35
Excellent	4	15

C. *Text-to-Speech Implementation*

Tensor flow environment was setup. Words and phrase was selected to test the accuracy of the system with default variables. As show in table II, fifteen phrases and words that contained different syllables were chosen. The output of the corresponding text were saved in a .wav file. The initial experimentation indicate that there is a promising result.

D. *Result and Discussion*

The means intelligent measure was used to ascertain the accuracy of the implemented prototype during the experiment. This shows the extent to which the synthetic speech is comprehensible. The fact that speech output produced depend on the default parameters.

As show in Table II, two speakers of Hausa language evaluated intelligibility of the speech output. The speakers were asked to write down what they hear as the listen to the speech output. The average number of syllables correctly listened to was computed against the number of syllables in the test to get the percentage of accuracy. The speakers average response accuracy was 87%.

TABLE II EXPERIMENT RESULT

Hausa Word/ Phrase	Syllables	Avg Correctly interpreted	Avg Intelligibility
Barka dai	2	1	50%
Zuma	1	1	100%
Ilimi boko	2	2	100%
Gari ya waye	3	2.5	83.3%
Dalibi mai kwazo da kokari	5	4.5	90%
Juma;a baba rana ce	4	3	75%
Iyaye mata sunu da hankali da tausayi	7	7	100%
Ilimi gishirin zaman duniya	4	3.5	87.5%
Yara manya gobe	3	3	100%
Ba’a fefe gora ranar tafiya	5	4	80%
Total	36	31.5	86.58%

V. CONCLUSION

This paper goal was to design a text corpus with the primary goal of building a Text to Speech system. The texts were collected, analyzed and evaluated. Afterwards, the filtered text were recorded. We were able to generate a speech dataset that comprises of 2,000 and 100,000 sentences and words respectively. The dataset was tested on an experimental TTS machine with a promising result. The weakness of the system is word base not phonemes based. Our future work will be to increase the size of the text corpus, and run the experiment on phonemes based.

Authorized licensed use limited to: Wikipedia. Downloaded on January 20,2025 at 09:39:54 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] P. Mukherjee, et. al, "Development of GUI for Text-to-Speech Recognition Using Natural Language Processing," IEEE, 2018

[2] J. Kanisha, G. Balakrishanan,Tdon, "Speech Transaction for Blinds Using Speech-Text-Speech Conversions", Communications in Computer and Information Science book series (CCIS), vol 131, part 1, pp.43–48, 2011

[3] C. S. T. Thus, T. Zin, "Implementation of Text to Speech Conversion," International Journal of Engineering Research & Technology, vol. 3(3) 2014, pp. 911–915.

[4] S. Patil, M. Phonde, S. Prajapati, S. Rane, A. Lahane, "Multilingual Speech and Text Recognition and Translation and Translation using Image,", International Journal of Engineering Research & Technology, vol. 5(4), pp. 85-87, 2016.

[5] A. Joshi, D. Chabbi, M. Suman, S. Kulkarni, "Text to Speech System for Kannada Language," International Conference on Communications and Signal Processing, 2015.

[6] A Naser, D. Aich, MD..Ruhul Amin, "Implementation of Subachan: Bengali Text to Speech Synthesis Software," 6th International Conference on Electrical and Computer Engineering ICECE, December 2010.

[7] M. Hamad, M. Hussain, "Arabic Text-To-Speech Synthesizer" IEEE Student Conference on Research and Development, 2011.

[8] Y-Ji. Sher, M-Chun Hsu, "Develop a HMM-based Tawanese Text-To-Speech System", 2nd International Conference on Software Technology and Engineering, 2010

[9] M Zeki, O. o. Khalifa, A. W. Naji, "Development of An Arabic Text-to_speech System", International Conference on Computer and Communication Engineering, May 2010

[10] H. U. Mullah, F. Pyrtuh, L. J Singh, "Development of an HMM-based Speech Synthesis System for Indian Englisg Language", International Symposium on Advance Computing and Communication, 2015

[11] M. L. Padmesh, P. S. Kumar, "Implementation of Viterbi Coder for Text to Speech Synthesis", International Conference on Computational Intelligence and Computing Research IEEE, 2015

[12] D. C. Tran, M. K. A. Ahamed Khan, S. Sridevi, "On the Training and Texting Data Preparation for End-to-End Text-to-Speech Application", 11th IEEE control and System Graduate Research Colloquium, August 2020

[13] M. Y. Arafar, S. Md. Jamirul Islam, Md. A. Siddiquee, A. Khan, M. R. A. Kotwal, M. N. Huda, "Speech Synthesis for Bangla Text to Speech Conversion", IEEE 2014

[14] M. Zhang, L. Zhao & H. Li, "Transfer Learning From Speech Synthesis to Voice Conversion With Non-Parallel Training Data," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 29, 2021.

[15] Y. Choi, Y. Jung & Y. Suh, "Learning to Maximize Speech Quality Directly Using MOS Prediction for Neural Text-to-Speech," IEEE Access, May 2022.

[16] T. Saeki, S. Takamicihi & H. Saruwatari, "Incremental Text-to-Speech Synthesis Using Pseudo Lookahead With Large Pretrained Language Model," IEEE Signal Processing Letters, Vol. 28, 2021.

[17] R. Liu, B. Sisman, G. Gao & H. Li, "Decoding Knowledge Transfer for Neural Text-to-Speech Training," IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 30, 2022.

[18] R. Kizito, W. S. Okello & S. Kagumire, "Design and Implementation of a Luganda Text Normalization Module for a Speech Synthesis Software Program," South African Institute of Electrical Engineers, Vol. 111(4), December, 2020.

Authorized licensed use limited to: Wikipedia. Downloaded on January 20,2025 at 09:39:54 UTC from IEEE Xplore. Restrictions apply.