JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Text To Speech Synthesizer

Shiburaj Pappu
Dept of Computer Science Engineering
Rizvi College Of Engineering
Mumbai,India
Shiburaj@eng.rizvi.edu.in

Nayna Dahatonde
Dept of Computer Science Engineering
Rizvi College Of Engineering
Mumbai,India
nayna@eng.rizvi.edu.in

Anwar Baig
Dept of Computer Science Eng.
Rizvi College Of Engineering
Mumbai,India
anwar433@eng.rizvi.edu.in

Mahima Patel
Dept of Computer Science Engineering
Rizvi College Of Engineering
Mumbai,India
mahimap@eng.rizvi.edu.in

kaif Shaikh
Dept of Computer Science Engineering
Rizvi College Of Engineering
Mumbai,India
kaifu9shaikh@gmail.com

Abstract— Text-To-Speech (TTS) conversion is a computer-based system that can read any text aloud, whether it was manually entered into the computer or scanned and sent to an OCR system. There are various systems that turn normal language text into speech in text to speech. The primary objectives are to investigate Optical Character Recognition in conjunction with speech synthesis technologies and to use Python to construct a cost-effective, user-friendly image to speech conversion system.We sometimes prefer to listen to information rather than read it.

While listening to the crucial file data, we can multitask. Python has a number of APIs for converting text to speech. The Google Text to Speech API is well-known and widely used.

Keywords— (TTS)text to speech, API, OCR(optical character recognition).

I. INTRODUCTION

The artificial fabrication of human speech is known as speech synthesis. A voice synthesiser is a computer system that can be implemented in software or hardware for this purpose. A text-to-speech (TTS) system translates text into speech; other systems render symbolic linguistic representations such as phonetic transcriptions into speech. Text-to-speech (TTS) is a convention that converts linguistic data or text into speech.

It is currently frequently utilised in audio reading systems for the blind. However, in recent years, text-to-speech conversion technology has expanded well beyond the disabled community, becoming a crucial complement to the rapidly expanding usage of digital voice storage. In Python, there are numerous APIs for converting text to speech. The Google Text to Speech API, often known as the gTTS API, is one of these APIs.

English, Hindi, Tamil, French, German, and many other languages are supported via the gTTS API. The speech can be given at either of the two audio speeds available: rapid or slow. However, as of the most recent update, changing the voice of the generated audio is not feasible.

TTS has a variety of applications in our daily lives: Telephony Automated telephone transactions (for example, banking operations), automated call centres for information services (for example, access to weather reports), and so on.

In-car devices such as the radio, air conditioning system, navigation system, cell phone (e.g., voice dialling), and others release automotive data.

II. LITERATURE SURVEY

International Journal of Engineering Research & Technology (IJERT) (Vol. 3 Issue 3, March - 2014**):**

MATLAB is used to transform an image to text and subsequently that text to speech in this project. E-text to speech conversion is also a success. Text from a word document, a Web page, or an e-Book can be read using this method, and synthesised speech can be generated using the computer's speakers. To convert an image to text, it must first be transformed into a grey image. By thresholding, a grey image is turned to a binary image, which is then converted to text in MATLAB. The OCR system is used in this project to recognise capital English characters A to Z and numbers 0 to 9. Each character is instantly recognised. In a notepad file, the identified character is saved as text. A text-to-speech conversion system is developed in this project.

- ***UIJRT | United International Journal for Research & Technology | Volume 02, Issue 03, 2021***

Text-to-Voice (TTS) synthesis can translate any input text into understandable and natural-sounding speech, allowing information to be transmitted from a machine to a person. It can be utilised by the handicapped and visually challenged as message readers, teaching assistants, communication aids, and learning aids.

Collecting text, preprocessing text, preparing phonetically balanced sentences, recording sentences, preparing an annotated speech database, and designing a prototype were all part of the process of developing Afaan Oromo voice synthesis. The system's performance was evaluated using the mean opinion score technique.

Six hundred sentences were used for training in this study, and ten arbitrary sentences were used to test the taught sentences. In terms of naturalness and intelligibility, we received 4.1 and 4.3 out of 5 scores, respectively, after training and testing our system. The prototype is trained and tested using a tenfold threshold method.

Tokuda et al. [31] create an English-language HMM speech synthesis system. For text analysis and feature extraction, such as contextual aspects, the authors employed festival speech tools. They used 524 sentences to train the model, with the voice stream sampled at a rate of 16 kHz. During speech synthesis, contextual cues were taken into account.

However, the authors did not conduct a quantitative study of the results, instead concluding that the delivered output of this synthesised speech is superior to that of other rule-based voice synthesisers, such as the formant-based approach. Ntsakoetal. [5] Using a hidden Markov model (HMM) speech synthesis technology, create a highly intelligible and natural-sounding voice synthesis system for the Xitsonga language.

They discovered that 7.7% of respondents thought the entire system was great, 38.9% thought it was good, 46 percent thought it was acceptable, and 7.7% thought it was terrible. This indicates that the authors obtained a 92.3 percent approval rating. This approach, on the other hand, can synthesis speech with only a few megabytes of training voice data. Finally, the speech parameters are used to generate synthetic speech.

A total of 500 sentences were utilised to train the model from a corpus of 11,670 sentences, and twenty sentences that were not included in the training dataset were used to assess the system's performance.

In general, despite the fact that these linked efforts have made a significant addition to the field, voice synthesis on the Afaan Oromo language has not yet been completely studied in the same way that other foreign languages have. According to the review, no work has been attempted to design a speech synthesiser for the Afaan Oromo language with the integration of non-standard words, which has advantages over other approaches such as requiring a smaller corpus for training, requiring very little memory, and being easily integrated into existing systems.

The ability to adjust voice qualities is another advantage of the HMM-based speech synthesis technology. As a result, the Hidden Markov model speech synthesis is the best and most significant approach to solving speech synthesis challenges such as naturalness, intelligibility, cost-effectiveness, expressivity, and producing average speech units, smooth to stable, and stores statistics rather than waveforms.

- **International Journal of Scientific & Engineering Research Volume 10, Issue 7, July-2019**

We have offered a survey of different TTS strategies as well as TTS difficulties in this paper. The text-to-speech communication is effective and efficient for users, resulting in high-quality speech. These approaches, formant synthesis, articulatory, and concatenative synthesis, are used to create the desired speech. Concatenative and formant synthesis are the most often employed approaches in today's systems.

Concatenative speech synthesis produces natural-sounding speech. An overview of a unit selection technique-based concatenative speech synthesis system. The unit length in the database has an impact on the quality of the synthesised speech. With longer units, the naturalness of the synthetic speech improves. However, more memory is required, and the number of units in the database grows significantly.

An additive synthesis and an acoustic model are used to create the synthetic speech. The method generates extremely understandable synthesised speech. The third fundamental method, articulatory synthesis, simulates a human's natural speech production process.

III. PROPOSED METHODOLOGY

TTS' primary goal is to turn any text into a waveform. Speech production involves the creation of an auditory waveform that corresponds to the text, as well as each of these units in the sequence. Text analysis, text normalisation, text processing, grapheme-to-phoneme conversion, and speech synthesis were all part of the process.

Text analysis examines the input text by breaking it down into words and sentences. Text normalisation is the process of transforming text into a form that may be spoken. It is the part of TTS that assigns phonetic transcription to each word. Grapheme to phoneme conversion is the process of assigning phonetic transcription to a word.

We define the common form model and various more models to help us understand how text to speech conversion by computer works. The Common Form model is as follows: A text analysis system decodes the text signal and exposes the form, and a speech synthesis system encodes this form as speech, are the two components of the common form model.

The first system involves resolving ambiguity in a noisy signal in order to produce a clean, unambiguous message; the second system involves encoding that message as a new, noisy, ambiguous, and redundant signal. We always read the words as they are contained in the text in the basic common form model; every word is read in the same way.

The key features of the model are

- Text analysis and speech synthesis are the two most basic procedures.
- Synthesis' duty is to generate a signal from this form.
- The process of transforming a written signal into a spoken signal is described in this model. Here, we transform text to speech in real time. In such models, the process is viewed as one of directly converting text into speech rather than unearthing a linguistic message from a written signal and then synthesising from it. The system is not separated into clear analysis and synthesis steps, for example.
- Signal to signal models are implemented as pipeline models, and the operation is similar to passing representation from one module to another. These systems are extremely modular. As a result, each module's function is specified as reading one sort of data and outputting another.

The phonetic analysis, often known as word analysis, is concerned with the phonetic structure of a word. Finally, sound is produced by symbolic verbal representation. TTS's flow diagram is shown in Figure 1.

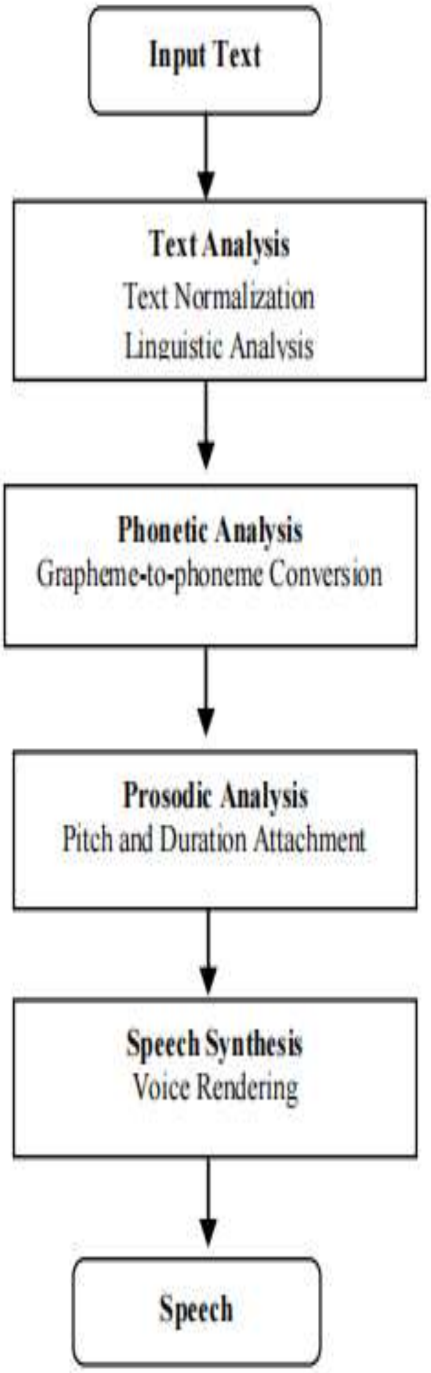

Fig 1: **Block diagram of proposed implementation**

Grapheme and phoneme form model: This method is similar to the comman form model in that it finds the grapheme form of the text input and converts it to a phoneme form for synthesis.

This model converts the grapheme form of the text input into phoneme form of speech synthesis, which is the actual pronunciation of each word in the input phrases. In contrast to the common form paradigm, words are not central to the representation.

RESULTS

This major project does not require any additional circuit setup as it mainly depend on programming approach.With the help of the Online Virtual platform we are developing our Project.

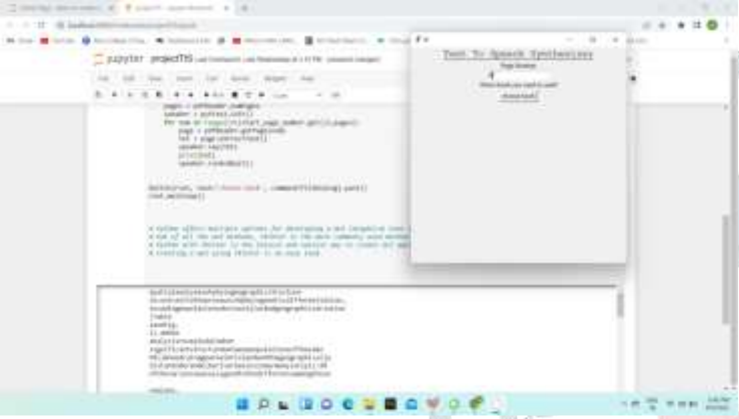

Fig 2. Text to speech

From the above image, we came to a conclusion that the accuracy of the project is 100%.

REFERENCES

[1]International Journal of Scientific & Engineering Research Volume 10, Issue 7, July-2019

[2] ***UIJRT | United International Journal for Research & Technology | Volume 02, Issue 03, 2021***

[3]International Journal of Engineering Research & Technology (IJERT) (Vol. 3 Issue 3, March - 2014)

[5] **IJRIT International Journal of Research in Information Technology, Volume 2, Issue 5, May 2014,**

[6] Assistive Technology by Courtney Lacomblez(2017)
https://www.readspeaker.com/blog/uses-text-speech-tts-anyway/

[7] F. Hinterleitner, S. Moller, T. H. Falk, and T. Polzehl, “Com- ¨ parison of Approaches for Instrumentally Predicting the Quality of Text-to-Speech Systems: Data from Blizzard Challenges 2008 and 2009,” in Proc. Blizzard Challenge Workshop, 2010

[7]J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan et al., “Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions,” in Proc. ICASSP, 2018.