2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP)

Smart Glasses: A Visual Assistant for Visually Impaired

Prabha M
Assistant Professor
Department of Information Technology
Velammal college of Engineering and Technology
Madurai, India
mpr@vcet.ac.in

Saraswathi P
Assistant Professor
Department of Information Technology
Velammal college of Engineering and Technology
Madurai, India
psw@vcet.ac.in

Hailly J
Student
Department of Information Technology
Velammal college of Engineering and Technology
Madurai, India
haillyshiela@gmail.com

Sindhuja C
Student
Department of Information Technology
Velammal college of Engineering and Technology
Madurai, India
csindhuja2002@gmail.com

Udhaya P
Student
Department of Information Technology
Velammal college of Engineering and Technology
Madurai, India
hariudhaya555@gmail.com

***Abstract* -- The blind people cannot read the text; they will suffer a lot in their day-to-day lives to handle this. Many techniques were introduced, but they didn’t provide better accuracy. The main aim of our project is to help the visually impaired by converting text into speech using Natural Language Processing (NLP). Text-to-speech (TTS) is an assistive technology that aids in digital text reading. The text is provided as an input to the processor, and as a result, audio will be produced by the speech synthesizer. Pre-processing, text-to- phoneme conversion, prosody generation, and text to-speech synthesiser conversion are all used in this research. The input text is split into chunks, and tokenized words are used for stemming which is used for text classification. The processed data is collected and stored in the speech synthesizer, and this synthesiser produces an audio file.**

Keyword -- Text-to-speech (TTS), Natural Language Processing (NLP), Text-to-phoneme, speech synthesizer.

I INTRODUCTION

The number of visually impaired people in this generation has increased over the past few decades, both locally and globally. The World Health Organization estimates that there are 255 million visually impaired persons worldwide, more than 36 million of whom are blind. Most people who are blind reside in low- and middle-income nations, where there is frequently inadequate access to facilities for eye care and rehabilitation. The various types of blindness, such as night blindness and colour blindness, affect different people in different ways. People who are partially blind are still able to function in society, but it becomes much more difficult for someone who has entirely lost their sight to survive in the modern world. Artificial intelligence has a subfield called machine learning, which can be thought of as a machine's capacity to imitate intelligent human behaviour. Here are a few illustrations: For those who are blind or visually impaired,

a. Text-to-speech (TTS) synthesis is a common use of machine learning (ML) since it gives them access to written information via speech. To produce speech from text, TTS system uses current technologies like deep learning methodologies.
b. Object Recognition: ML algorithms may be used to create computer vision systems that can identify barriers and objects, giving blind people knowledge of their surroundings. To recognize objects, these systems often employ convolutional neural networks (CNNs) that have been trained on massive image datasets.
c. Speech Navigation: For people who are blind or visually challenged, voice navigation systems can also be created using ML algorithms. To give spoken instructions and react to voice requests, these systems employ TTS synthesis and speech recognition algorithms.
d. Emotion Recognition: By using ML algorithms to create systems that can understand and react to

979-8-3503-4842-2/23/$31.00 ©2023 IEEE

emotional expressions in speech, visually impaired people will be able to communicate more naturally and expressively.

The processes involved in the suggested techniques like RNN and CNN using TTS synthesis for blind people. We proposed to use Natural Language Processing (NLP) algorithms for TTS synthesis is as follows:

The first stage is the process of converting raw text input into a form, which is referred to as "pre- processing". The second stage is the process of converting written text to phonetic sequence using text-to-grapheme conversion. The third stage, prosody generation, makes sure that the speech produced by the TTS system sounds natural and is understandable. Voice synthesis is the last stage, it produces the speech signal from the prosodic parameters and phoneme sequence using a TTS system.

II LITERATURE SURVEY

An end-to-end text-to-speech system with efficient emotion transfer aims to strengthen the emotional expression of the text-to-audio detection system. Emotion encoder and E2E TTS mode are the two components that derive expressiveness condition vectors using text input [2].

Exemplar-based emotional speech synthesis uses a database of recorded speech samples to generate speech with a particular emotion. This method was used to collect and pre-process the speech exemplars, including techniques for speaker normalization, emotion annotation, concatenative synthesis, statistical parametric synthesis, and generative models [8].

Incremental TTS synthesis is the process that generates audio on a word-by-word or sentence-by- sentence basis, as opposed to generating speech for the entire text at once. The use of large pre-trained language model has been proposed as a method for incremental TTS synthesis offers improved speech quality and a more natural flow [6].

Audio enhancement technique is used to reduce the effect of noise and reverberation in audio signals that improves the quality and intelligibility of generated speech [5].

The use of sound recognition technology for real- time labelling and course transcription in the learning platform provides real-time captions of spoken content to students, making it more accessible to those with hearing impairments or difficulties. Speech recognition systems are integrated into the classroom to transcribe the spoken words of the instructor in real-time [11].

Based on deep neural networks for time-variant linear transformations between the source and target speaker's voice which is used to convert the speech-to-speech synthesis for preserving the linguistic content of the audio. [12].

The goal of speech conversion for murmuring audio synthesis is to convert normal audio into murmuring audio while preserving the linguistic content of the speech. Whispered speech has different acoustic characteristics compared to normal speech. The deep learning-based voice conversion methods have the potential to produce high-quality whispered speech synthesis [13].

Expressive TTS training is intended to produce audio that not only conveys the linguistic content of the text but also captures the emotional and stylistic information, such as prosody and rhythm [14].

A Comparison between Straight, Glottal, and Sinusoidal Vocoder in Statistical Parametric Speech Synthesis is a research area that focuses on comparing different vocoding techniques which are used to synthesize speech [15].

A Machine Audio Chain Approach for Dynamically Adaptive Lombard TTS in Static and Dynamic Noise platforms is a research area that focuses on using a machine speech chain approach for text-to-speech (TTS) synthesis in noisy and dynamic environments.
[16].

A Textual-to-Audio Pipeline, based on the Test Results, and Initial Results of Fine-Tuning for Preschool Voice The goal of synthesis is to create text-to-audio synthesis technology for children's voices. This information is then utilised to generate acoustic signals, which are further processed to produce the final speech output [3].

An Audio-Visual System for Object-Based Audio: From Recording to Listening is to develop a multimedia system for creating, processing, and delivering object-based audio [17].

The development of a voice command programme text normalisation module programme in the Luganda language is referred to as a Luganda Textual Standardization Component for a Voice Generation Software Program. the speech synthesis software to generate more accurate and natural-sounding speech output [9].

Multi-Scale Feelings Transference, Forecasting, and Management of Emotionally Voice recognition delivers good performance, allowing emotive voice recognition by transferring emotion from sound or estimating emotion from only the input sequence [12].

The goal of decrypting transfer of knowledge for neural message training is to enhance the performance and efficacy of TTS systems through knowledge transfer methods such as transfer learning and fine-tuning, as well as to assess the effectiveness of these techniques in terms of speech quality, naturalness, and efficiency [1].

End-to-End Psychological Voice Composition Using an Appropriate Style Phrase Load Method of Control, focusing on techniques for controlling the emotional expression of synthesised speech using style tokens [4].

A Local Attention Mechanism-Based Word Consolidation Algorithm for Voice Synthesis, which focuses on techniques for normalising text input to enhance the accuracy of sound generation, with a specific emphasis on the use of local attention mechanisms [18].

III PROPOSED METHODOLOGY

Pre-processing:

Pre-processing in text-to-speech (TTS) refers to the process of transforming raw text input into a form that can be processed and synthesized by a TTS system. As shown in Fig. The pre-processing step is crucial in ensuring that the TTS system produces high-quality, natural-sounding speech. The processing steps may involve various tasks, such as

a. Tokenize the text into words:

The text is first tokenized, or broken down into smaller units, such as words, sentences, or phrases, and then each unit is synthesized into speech.

b. Part-of-speech tagging for each word:

It is the process of identifying every word in the text with its appropriate portion of speech, such as a noun, verb, adjective, etc. The parts of speech (POS) tags help the system generate more accurate and natural- sounding speech.

c. Prosodic Analysis:

"Prosody" refers to the rhythm, stress, and intonation of speech. It involves analysing the text to be spoken and determining the appropriate prosodic features, such as stress, rhythm, and intonation, to be used when generating speech. It helps the system produce speech that is more natural and easier to understand.

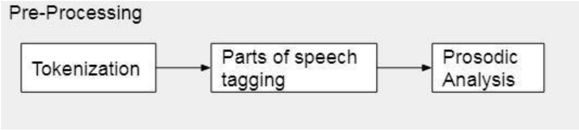

Fig. 1. Pre-Processing

Text-to-Phoneme Conversion:

It is the process of generating speech from text, especially in the context of a textual to sound system for the blind. The purpose of word to morpheme translation is to transform written words into a phonetic pattern.

a. Grapheme to phoneme:

The function of morpheme to phonetic conversion is to transform written words into a pronunciation pattern that accurately captures the text's pronunciation. It is typically done using a combination of rule-based and machine-learning approaches.

b. Combine phoneme:

Combine the phonemes for each word to form a phoneme sequence that accurately represents the pronunciation of the text. It helps to ensure that the speech generated by the TTS system is accurate and natural sounding.

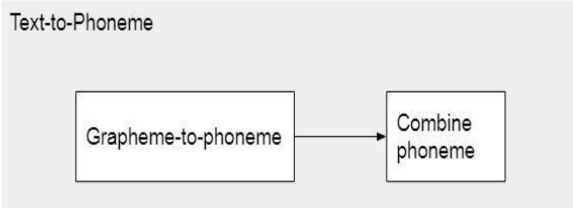

Fig. 2. Text-to-Phoneme (T2P)

Prosody Generation:

Prosody generation ensures that the speech generated by the TTS system is natural-sounding and easy to understand. Prosody refers to the rhythm, stress, and intonation of speech and plays a crucial role in conveying the meaning and emotion of spoken language. In the context of TTS for the blind, prosody generation involves using the information contained in the text and the phoneme sequence to control the rhythm, stress, and intonation of the speech generated by the TTS system.

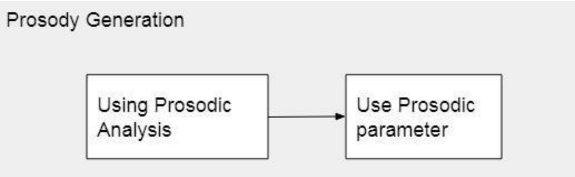

Fig. 3. Prosody Generation

Speech Synthesis:

Speech synthesis is applied to generate the utterance parameters' audio signal and phoneme sequence and save the speech signal as an audio file.

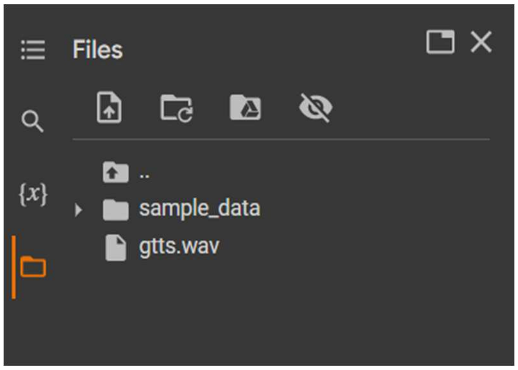

Fig. 4. Audio format of wav file

Fig. 5. Text to speech in Audio format

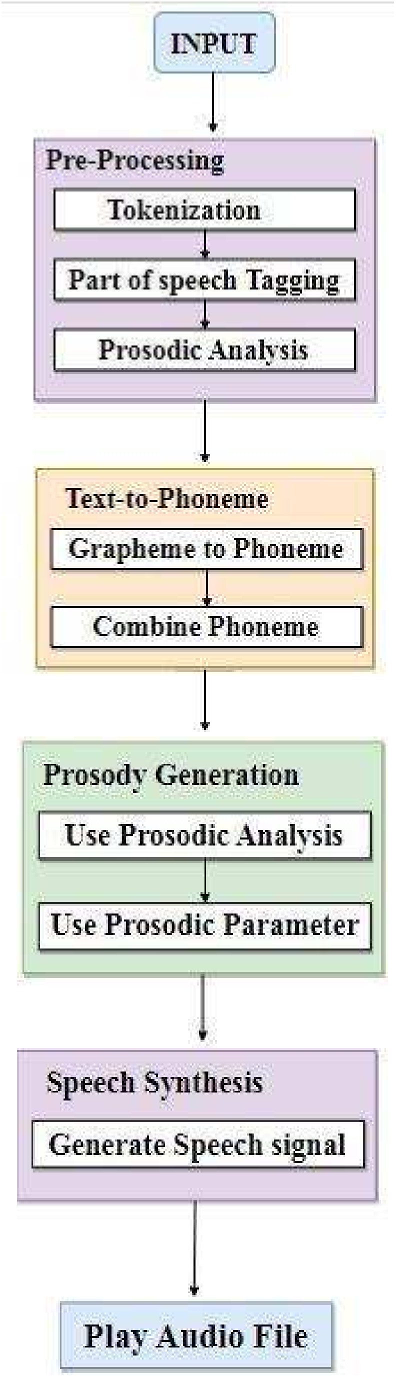

Fig. 6. Block Diagram of Proposed Methodology

IV RESULTS AND DISCUSSION

Text-to-speech (TTS) conversion is accomplished through pre-processing, text-to-phoneme conversion, prosody generation, and final output from the speech synthesizer. Pre-processing helps bring the text into a form that is predictable and analysable. During pre- processing, tokenization is performed to break non- relational and language processing content into information snippets that can be regarded as separate parts. The prosodic analysis is used to identify features such as stress, rhythms, and intonation. The intonation is used to analyse the variance in the voice and determine whether the range is high or low. After completing the pre-processing, the processed data is passed to the text-to-phoneme processor. In this text-to-phoneme processor, two processes are performed. At first, the grapheme-to-phoneme process is initiated; in this conversion, the input text (letters) is processed, and the pronunciation for these letters is produced as a result. After this step, all the chunks of the phoneme are combined and stored in a separate dataset. The grapheme-to-phoneme process of Audio to Text and Automated Voice Recognition depend heavily on it.

The voice is generated via a voice recognition signal from the prosodic parameter and phoneme sequence. The processed and final speech signal is saved as an audio file.

A scatterplot is essentially focused with the coordinates X and Y. The coordinates are applied to classify the information that is provided as input. Matplotlib is used to draw as well.

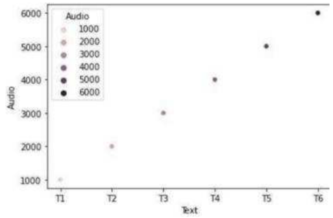

Fig. 7. Scatter Plot of audio recognition

The arrangement of the coordinates in swarm plot is customized according to the requirements. This methodology reduces data overlap. The Seaborn libraries support swarm plot, as well as the graphs were obtained by plotting using matplotlib.

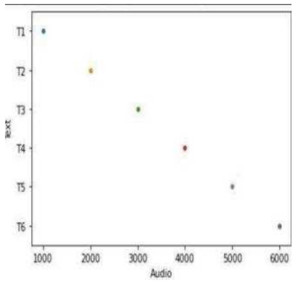

Fig. 8. Swarm plot of audio recognition

The input text is evaluated using natural language processing (NLP), and the resulting data is stored in a csv file with both the XY parameters representing the text and audio fields. If the wavelength or amplitude of the voice is raised, the graph increases in proportion to the length of the sentence.

The accuracy of existing solutions is 83.7%. The accuracy obtained from text-to-speech (TTS) is 97.3%. Our proposed system's accuracy is high due to the use of libraries such as NumPy, TensorFlow, Librosa, and gTTs.

V CONCLUSION

This project is implemented by using a text-to-speech algorithm using NLP. There are several tools available for converting text to speech. The reason for choosing TTS is its better accuracy as compared to others. The accuracy provided by TTS is 97.3%, which is higher than the other proposed solutions. The accuracy is obtained by using some libraries, such as Librosa, NumPy, TensorFlow, and gTTS. The TTS involves pre-processing, prosody generation, text-to-phoneme conversion, and a speech synthesizer, which will have several modules that will convert the text to audio. The scope of the project is to help blind people.

VI REFERENCES

[1] R. Liu, B. Sisman, G. Gao and H. Li, "Decoding Knowledge Transfer for Neural Text-to-Speech Training," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1789-1802, 2022,doi: 10.1109/TASLP.2022.3171974.

[2] O. Kwon, I. Jang, C. Ahn and H. -G. Kang, "An Effective Style Token Weight Control Technique for End-to-End Emotional Speech Synthesis," in IEEE Signal Processing Letters, vol. 26, no. 9, pp. 1383- 1387, Sept. 2019, doi: 10.1109/LSP.2019.2931673.

[3] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran and H. Cucu,"A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis," in IEEE Access, vol. 10, pp. 47628-47642,2022, doi: 10.1109/ACCESS.2022.3170836.

[4] Y. -S. Joo, H. Bae, Y. -I. Kim, H. -Y. Cho and H. -G. Kang, "Effective Emotion Transplantation in an End-to-End Text-to- Speech System," in IEEE Access, vol. 8, pp. 161713-161719,2020, doi: 10.1109/ACCESS.2020.3021758.

[5] C. Valentini-Botinhao and J. Yamagishi, "Speech Enhancement of Noisy and Reverberant Speech for Text-to-Speech," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 8, pp. 420-1433, Aug. 2018, doi: 10.1109/TASLP.2018.2828980.

[6] T. Saeki, S. Takamichi and H. Saruwatari, "Incremental Text-to-Speech Synthesis Using Pseudo Lookahead with Large Pretrained Language Model," in IEEE Signal Processing Letters, vol. 28, pp. 857-861,2021, doi: 10.1109/LSP.2021.3073869.

[7] Y. Lei, S. Yang, X. Wang and L. Xie, "MsEmoTTS: Multi-Scale Emotion Transfer, Prediction, and Control for Emotional Speech Synthesis," in IEEE/ACM Transactions on Audio, Speech, and LanguageProcessing,vol.30,pp853864,2022,doi: 10.1109/TASLP.2022.3145293.

[8] X. Wu et al., "Exemplar-Based Emotive Speech Synthesis," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 874-886, 2021, doi: 10.1109/TASLP.2021.3052688.

[9] R. Kizito, W. S. Okello and S. Kagumire, "Design and implementation of a Luganda text normalization module for a speech synthesis software program," in SAIEE Africa Research Journal, vol. 111,no. 4, pp. 149-154,Dec. 2020, doi: 10.23919/SAIEE.2020.9194384.

[10] Y. Choi, Y. Jung, Y. Suh and H. Kim, "Learning to Maximize Speech Quality Directly Using MOS Prediction for Neural Text-to-Speech," in IEEE Access, vol. 10, pp. 52621-52629, 2022, doi: 10.1109/ACCESS.2022.3175810.

[11] R. Ranchal et al., "Using speech recognition for real-time captioning and lecture transcription in the classroom," in IEEE Transactions on Learning Technologies, vol. 6, no. 4, pp. 299-311, Oct.-Dec. 2013, doi: 10.1109/TLT.2013.21.

[12] G. Kotani, D. Saito and N. Minematsu, “Voice Conversion Based on Deep Neural Networks for Time-Variant Linear Transformations," in IEEE/ACM Transactions on Audio, Speech, and Language Processing,vol.30,pp.2981-2992,2022,doi:10.1109/ TASLP.2022.3205755.

[13] M. Cotescu, T. Drugman, G. Huybrechts, J. Lorenzo-Trueba and A. Moinet, "Voice Conversion for Whispered Speech Synthesis," in IEEE Signal Processing Letters, vol. 27, pp.186-190, 2020, doi: 10.1109/LSP.2019.2961213.

[14] R. Liu, B. Sisman, G. Gao and H. Li, "Expressive TTS Training With Frame and Style Reconstruction Loss," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 29, pp. 1806-1818, 2021, doi: 10.1109/TASLP.2021.3076369.

[15] M. Airaksinen, L. Juvela, B. Bollepalli, J. Yamagishi and P. Alku, "A Comparison Between STRAIGHT, Glottal, and Sinusoidal Vocoding in Statistical Parametric Speech Synthesis," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 9, pp. 16581670, Sept.2018, doi:10.1109/TASLP.2018.283572.

[16] S. Novitasari, S. Sakti and S. Nakamura, "A Machine Speech Chain Approach for Dynamically Adaptive Lombard TTS in Static and Dynamic Noise Environments," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 2673-2688, 2022, doi: 10.1109/TASLP.2022.3196879.

[17] P. Coleman et al., "An Audio-Visual System for Object-Based Audio: From Recording to Listening," in IEEE Transactions on Multimedia, vol. 20, no. 8, pp. 1919-1931, Aug. 2018, doi: 10.1109/TMM.2018.2794780.

[18] L. Huang, S. Zhuang and K. Wang, "A Text Normalization Method for Speech Synthesis. Based on Local Attention Mechanism," Access,vol.8,pp.3620236209,2020,doi:10.1109/ACCESS.2020.2974674.