iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.51217

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Text To Speech with Custom Voice

Nigam R Poojary[1], K H Ashish[2], Pavan[3], Pradeep[4], Sunitha N V[5]
[1, 2, 3, 4]*Student,* [5]*Assistant Professor, Dept of Computer Science and Engineering, Mangalore Institute of Technology and Engineering, Moodabidri*

Abstract: ***The Text to Speech with Custom Voice system described in this work has vast applicability in numerous industries, including entertainment, education, and accessibility. The proposed text-to-speech (TTS) system is capable of generating speech audio in custom voices, even those not included in the training data. The system comprises a speaker encoder, a synthesizer, and a WaveRNN vocoder. Multiple speakers from a dataset of clean speech without transcripts are used to train the speaker encoder for a speaker verification process. The reference speech of the target speaker is used to create a fixed-dimensional embedding vector. Using the speaker embedding, the synthesizer network based on Tacotron2 creates a mel spectrogram from text, and the WaveRNN vocoder transforms the mel spectrogram into time-domain waveform samples. These waveform samples are converted to audio, which is the output of our work. The adaptable modular design enables external users to quickly integrate the Text to Speech with Custom Voice system into their products. Additionally, users can edit specific modules and pipeline phases in this work without changing the source code. To achieve the best performance, the speaker encoder, synthesizer, and vocoder must be trained on a variety of speaker datasets.***

Keywords: Text-to-Speech, Tacotron2, WaveRNN, mel spectrogram, encoder, synthesizer, vocoder.

I. INTRODUCTION

The text-to-speech (TTS) technology has been around for a while and has made significant advancements recently. It involves creating human-sounding speech from written text, enabling the translation of any written content into speech. TTS technology has a variety of uses, including generating voiceovers for videos, providing audio output for those who are blind, and improving the accessibility of digital content. TTS technology has advanced from the earliest rule-based systems that produced speech using pre-recorded voice samples. These systems lacked adaptability to accommodate various languages and accents, which limited their capacity to produce speech that sounded natural. With the introduction of machine learning algorithms and the availability of massive datasets of speech recordings, TTS systems have advanced significantly and can now produce speech that is similar to that of a human. Currently, text-to-speech (TTS) technology finds application in various domains such as mobile devices, personal assistants, and navigation systems, among others. Developers can easily add speech synthesis to their applications without the requirement for custom voice cloning thanks to the availability of pre-built TTS voices. However, there are still circumstances where a distinctive voice is preferred, such as in branding or marketing, where it might help in giving a product or service its own personality. Voice cloning is one method for developing a Text To Speech with a custom voice system. The way we engage with digital assistants, video game characters, and even our own personal devices could be completely changed by this technology. Text To Speech with Custom Voice enables the immediate synthesis of speech in a specific voice, making it perfect for live performances or virtual assistants. Deep learning is used in voice cloning to train a neural network on recordings of a dataset. By recognizing patterns and connections between the many aspects of speech, such as pitch, intonation, and pronunciation, the neural network learns to produce new speech that mimics the original voice. The first step in voice cloning involves gathering a dataset containing a significant number of recordings of the speakers. The dataset is then pre-processed to extract pertinent features, like spectrograms or Mel frequency cepstral coefficients (MFCCs), after it has been gathered. A deep learning model is then trained to predict the associated audio output using these features as inputs. During the training process, the neural network's parameters are adjusted to minimize the difference between the actual audio output and the expected audio output. Backpropagation, an iterative method for accomplishing this, compares the network's output to the desired output and modifies the network as necessary.

II. NEED OF THE STUDY

The growing desire for more individualized and natural audio content, as well as the development of a system that can clone voices without requiring a lot of training data or retraining the model, have led to the necessity for research into text-to-speech technology with customized voice cloning. At present, the process of training a deep neural network for voice cloning requires several hours of recorded speech from a single speaker, which can be costly and time intensive.

We can decrease the quantity of training data needed and boost the effectiveness of the voice cloning process by creating a system that can duplicate voices using small samples of reference speech. With the use of this technology, users' experiences can be improved and more people can access digital information. In certain cases where obtaining large amounts of training data is not feasible, the ability to perform voice cloning using only a small amount of data can be beneficial. For instance, this may be useful when the speaker is no longer available for recording or has passed away. Therefore, we may enhance the effectiveness and accuracy of voice cloning systems, making it simpler to create personalized voices for a variety of applications, by investigating new deep learning approaches and designing novel algorithms. The field of text-to-speech technology with personalized voice cloning desperately needs these studies to advance.

III. LITERATURE REVIEW

A. TACTRON: Towards end-to-end speech synthesis [1]

In this study, the authors presented a new text-to-speech model called Tacotron, which can generate speech solely from characters. Tacotron1 is based on the sequence-to-sequence (seq2seq) model with attention, and it can be trained from scratch using pre-existing text and audio pairs with random initialization.

The model uses several methods to enhance the basic seq2seq model's functionality to generate a raw spectrogram from input characters. The authors outline numerous important methods for maximizing the performance of the sequence-to-sequence architecture in this difficult endeavor. Prior to sending character embeddings through the encoder, they employ a pre-net to forecast attention weights and lower the dimensionality of the embeddings. They additionally enhance the output spectrogram using a post-processing network.

The use of a neural vocoder to transform the generated auditory characteristics into a waveform is one of the TACTRON paper's major contributions. This method decouples the creation of acoustic features from the synthesis of the final waveform, allowing for more freedom in the production of speech.

One of the key contributions of the TACTRON paper is the use of a neural vocoder to convert the generated acoustic features into a waveform. This approach allows for greater flexibility in the generation of speech, as it decouples the generation of acoustic features from the synthesis of the final waveform.

B. Efficient Neural Audio Synthesis [2]

This paper outlines a collection of general techniques for reducing sampling time for audio synthesis while maintaining high output quality. The authors define WaveRNN as a text-to-speech synthesis technique. It is a single-layer recurrent neural network for audio generation. Compared to a GPU, the network can output 24 kHz, 16-bit audio more quickly. Second, they reduce the WaveRNN's weights using a weight-pruning approach. For a given set of parameters, they find that large sparse networks outperform small dense networks, and this association holds for sparsity levels higher than 96%. Finally, they suggest a new generation method based on subscale that allows for the simultaneous synthesis of numerous samples by folding a lengthy sequence into a group of shorter sequences. The Subscale WaveRNN provides an orthogonal strategy for improving sampling efficiency and can generate 16 samples each step without sacrificing quality.

C. Transfer Learning from Speaker Verification to Multi-speaker Text-To-Speech Synthesis [3]

This study suggests a new approach for transfer learning-based multi-speaker text-to-speech synthesis. The challenge of producing natural-sounding speech from the text in the voices of various speakers is discussed in the paper. An encoder and a decoder for speakers are the foundation of the suggested technique.

A neural network that has been trained on a sizable corpus of speaker verification data can be used to extract speaker-specific features from voice signals using the speaker encoder. The decoder, a neural network, creates the appropriate speech waveform using the text and speaker-specific information as input.

Mean opinion scores (MOS), which gauge how naturally the generated speech is considered to sound and how similar it is to actual speakers, were used to evaluate the suggested method. The proposed model enhances the stability of the synthesizer for multi-speaker synthesis in open-set scenarios. However, there is a noticeable disparity in speaker similarity between the synthesized speech and the natural speech of the speaker. The paper introduces transfer learning as a crucial element, where a speaker verification model is adjusted to cater to multi-speaker text-to-speech synthesis. This approach allows the model to capture the unique characteristics of each speaker's voice, even with limited training data.

D. Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System [4]

In this study deep neural networks are used to investigate speakers and languages. The paper covers the issue of precisely identifying the speaker or language of an audio stream, which is crucial for a variety of applications, including speech recognition, forensics, and security. The authors suggest a brand-new method for recognizing speakers and languages based on an end-to-end deep neural network that employs an encoding layer and a cutting-edge loss function. A neural network used in the encoding layer learns to extract high-level properties that are unique to the speaker or language from the raw audio signal as its input. In order to increase the separation between different speakers or languages in the learned feature space and increase recognition system accuracy, a novel loss function was developed. The suggested method was tested using different benchmark datasets, and the findings demonstrated that it performed better in terms of accuracy and robustness to noise and other sources of variability than current state-of-the-art methods. The authors also carried out an ablation study to evaluate the contribution of various system parts, such as the loss function and encoding layer.

E. Char2Wav: End-to-end Speech Synthesis [5]

The paper "Char2Wav: End-to-end Speech Synthesis" suggests a novel approach for speech synthesis based on end-to-end neural network architecture. The study addresses the issue of producing natural-sounding speech from text, which is crucial for a variety of applications, including speech therapy, virtual assistants, and audiobooks. The suggested method uses a recurrent neural network (RNN) to produce the matching speech waveform from character-level text input. To learn the mapping between text and speech, a sizable corpus of speech data is used to train the network. The authors also suggest a brand-new training objective that combines a perceptual loss based on the Mel-scale spectrogram distance with the conventional mean squared error (MSE) loss. The report provides a thorough investigation of the suggested approach, including an analysis of the network's various components through ablation research. The results demonstrate that the suggested approach performs better in terms of naturalness and resemblance to the target speaker than current state-of-the-art approaches. The suggested strategy is also tested on a number of benchmark datasets by the authors, who demonstrate that it outperforms current methods in terms of both objective and subjective measures, such as mean opinion scores (MOS) and log-spectral distances.

F. High-Fidelity and Low-Latency Universal Neural Vocoder Based on Multiband WaveRNN with Data-Driven Linear Prediction for Discrete Waveform Modelling [6]

The research paper suggests a brand-new neural vocoder architecture based on a multiband WaveRNN with a data-driven linear prediction for discrete waveform modeling. The purpose of this study is to develop neural vocoders, which are used to create high-quality audio waveforms from low-dimensional information like mel-spectrograms or linear predictive coding (LPC) coefficients.Two key parts make up the suggested neural vocoder: a multiband WaveRNN and a data-driven linear prediction (DDLP) module. The multiband WaveRNN design separates the audio signal into various frequency bands and uses a WaveRNN to treat each band independently. This strategy keeps the model's high audio quality while lowering its computing cost. The residual signal between the original audio waveform and the WaveRNN output is modeled using the DDLP module, significantly enhancing the audio quality. The authors assess how well their suggested neural vocoder performs on a variety of audio tasks, such as voice synthesis and music creation. They evaluate the performance of their model against a number of cutting-edge neural vocoders, such as WaveRNN and WaveNet. The authors discover that their suggested model outperforms competing models in terms of computing efficiency while delivering state-of-the-art audio quality. The impact of several hyperparameters on the effectiveness of the authors' model is also studied. The researchers discover that employing more frequency bands and WaveRNN layers results in better audio quality. They also point out that adding more frequency bands may result in higher computing costs.

Finally, by training their model on numerous datasets and testing it on diverse audio tasks, the authors show how generalizable their approach is. They discover that their model can be trained on many datasets with variable audio properties and yet perform at a cutting-edge level. This suggests that their model has the potential to function as a general-purpose neural vocoder that can produce high-quality audio for a variety of audio tasks.

G. Combination of deep speaker embeddings for diarisation [7]

An approach for encoding speaker identity in speaker recognition tasks using deep speaker embeddings is suggested in this paper. In order to predict speaker identification labels on the training data, the study explains how deep speaker embeddings are learned discriminatively. The embeddings should be robust to non-speaker acoustic fluctuation while capturing the variances between all potential speakers, recording the various acoustic elements that constitute a speaker's identity.

The use of deep speaker embeddings for diarization and verification tasks is also covered in the study. The deep learning method for speaker detection presented in this research involves constructing low-dimensional speaker embeddings from unprocessed audio data. The study shows that the suggested method outperforms conventional speaker recognition systems that rely on manually created features, such as mel-frequency cepstral coefficients (MFCCs), and that it achieves state-of-the-art outcomes on a number of benchmark datasets for speaker recognition, diarization, and verification. The study demonstrates other applications for the learned embeddings, including speaker diarization, which entails counting the number of speakers in a recording and grouping the portions of the recording that correspond to each speaker.

H. A comparative study of feature extraction techniques for speech emotion recognition [8]

This paper presents a comparative study of various feature extraction techniques for speech emotion recognition. The performance of several feature types, including prosodic features, spectral features, and higher-level features based on deep neural networks, is examined by the authors. They assess the impact of dimensionality reduction and feature selection approaches on the effectiveness of the emotion recognition system. The study used speech recordings of actors delivering scripted and spontaneous dialogues with emotional content from the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database. Support vector machines, k-nearest neighbors, and random forests were just a few of the classifiers that the authors trained and evaluated using the retrieved features. With an accuracy of 68.5% for four-class emotion recognition, the study's findings demonstrated that deep neural network features surpassed all other feature categories. However, the authors point out that the size of the training set had a significant impact on how well the deep neural network features performed, therefore learning these features efficiently may not be possible with smaller training sets. The study also discovered that the effectiveness of the emotion identification system was significantly impacted by feature selection and dimensionality reduction strategies. The authors discovered that principle component analysis (PCA) was the most efficient dimensionality reduction technique, whereas feature selection utilizing mutual information produced the highest result.

IV. METHODOLOGY

The system consists of three neural networks that are trained independently. The first network is responsible for encoding the speaker's voice, which produces a fixed-dimensional vector from the speech signal. The second network is a synthesizer that can generate mel spectrograms using inputs like phonemes or graphemes. Finally, the third network uses an autoregressive WaveRNN vocoder to convert the generated spectrogram into time domain waves. These three networks work together to produce synthetic speech.

A. Speaker encoder

The role of the speaker encoder network is to take in the audio of a specific speaker and transform it into a low-dimensional vector embedding that captures the distinct features of that speaker's voice. To achieve this, the network is trained on a dataset containing hundreds of speakers to perform speaker verification. The synthesis network then utilizes the embeddings by taking a reference speech signal from the target speaker. To accomplish speaker verification, a scalable and precise neural network architecture is used. The network maps a sequence of log-mel spectrogram frames extracted from any voice utterance of any duration into a fixed-dimensional embedding vector, referred to as a d-vector. The network does not rely on transcripts but instead uses a training dataset of 1.6-second-long spoken audio segments, each with a label for speaker identification.

The encoder model consists of a 3-layer Bidirectional LSTM with 128 hidden nodes and 64 fully connected units. The final embedding is produced by the output of the last layer, which has undergone L2-normalization. The model takes in 40-channel log-mel spectrograms calculated from 1.6-second voice utterances as input, with a 10ms step and a 25ms window width.

B. Synthesizer

To generate the Mel spectrogram from text, we employ a sequence-to-sequence neural network based on Tacotron 2, which is conditioned on the speaker embedding. At each time step, the speaker embedding is concatenated with the synthesizer encoder output to condition the output and match the speaker's voice.

The synthesizer is trained using pairs of target audio and text transcripts. We map the text to a series of phonemes at the input to facilitate quicker convergence and better pronunciation of uncommon words and proper nouns. During training, an already-trained speaker encoder extracts a speaker embedding from the target audio in transfer learning mode. The target mel spectrograms for the synthesizer contain more features than those used for the speaker encoder.

They have 80 channels and are computed with a 50 ms window and a 12.5 ms step. Our approach does not filter input texts for pronunciation, but we apply a few cleaning techniques such as replacing acronyms and numerals with their full textual forms, converting all letters to lowercase, and ensuring all characters are in ASCII format.

C. Vocoder

The system employs WaveRNN as a vocoder, which utilizes the Mel spectrograms generated by the synthesis network to produce time-domain audio waveforms via autoregressive synthesis. The synthesizer network is trained to capture all the necessary details for high-quality synthesis of various voices in the form of Mel spectrograms, allowing for the straightforward creation of a multispeaker vocoder by training on data from multiple speakers.

WaveRNN replaces all 60 of WaveNet's convolutions with a single GRU layer design. The Mel spectrogram and associated waveform are segmented into the same number of sections for each training step. An upsampling network is utilized to adjust the length of the target waveform based on the Mel spectrogram. To provide conditioning characteristics for the layers during the translation of the Mel spectrogram into a waveform, a resnet-like model also takes the spectrogram as input. The resulting vector is iterated over to fit the length of the waveform segment, and the upsampled spectrogram and the waveform segment from the previous time step are concatenated with each of the four sections of the conditioning vector. One of the primary advantages of WaveRNN is its ability to produce high-quality, authentic-sounding audio samples with minimal parameters. The Mel spectrogram predicted by the synthesizer network captures all the necessary information for the high-quality synthesis of a wide range of voices, allowing for the straightforward creation of a multispeaker vocoder by training on data from several speakers.

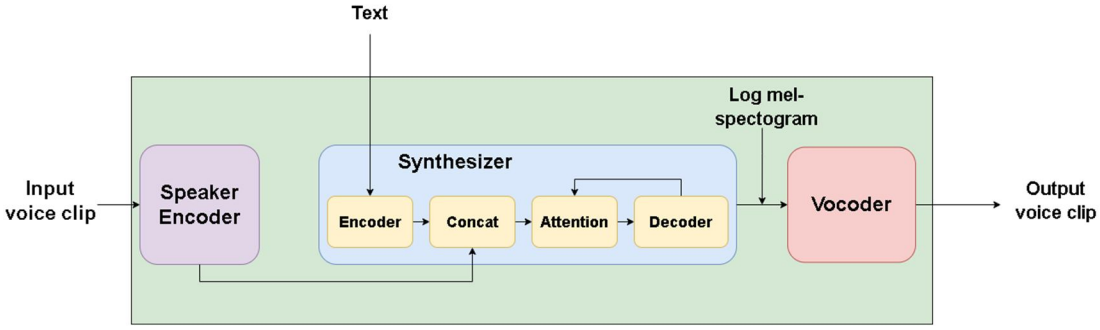


Fig.4.1 Components of Text-To-Speech With Custom Voice System

The figure 4.1 shows the components and workflow of the TTS system. The process of creating speech in a new speaker's voice is a multi-step procedure that employs advanced deep neural network models. The initial stage is to use the speaker encoder model, commonly known as Dvec, to convert the distinctive qualities of the target speaker's voice into a fixed-length vector representation. This involves feeding the Dvec model with numerous hours of professionally recorded speech from the target speaker, allowing the model to extract their unique voice features. The resulting vector representation is then utilized to adjust the synthesized speech to replicate the voice characteristics of the target speaker.

The subsequent step in the speech generation process involves using the Tacotron 2 synthesizer model to produce speech from a given text input. This model accepts a sequence of text tokens as input and generates a mel-spectrogram representation of the corresponding speech signal. Tacotron 2 is trained on a dataset of text-to-speech pairs, where each text input is paired with a corresponding speech signal.

The generated mel-spectrogram is then processed by the Wave RNN vocoder model, which converts it into a raw audio signal that can be played back as speech. The Wave RNN model is trained on a dataset of mel-spectrograms and their corresponding audio signals, allowing it to generate audio signals that closely match the input mel-spectrogram.

The final stage involves applying the Dvec model to modify the Tacotron 2's mel-spectrogram so that it matches the target speaker's distinct voice characteristics. This is done by calculating a distance metric between the target speaker's Dvec vector and the speaker's original Dvec vector from Tacotron 2's training. This distance measure is subsequently reduced in the mel-spectrogram, resulting in speech that resembles the target speaker's voice. The naturalness and variety of synthesised speech could be greatly improved by the use of cutting-edge deep neural network models to capture voice features, generate speech from text, and convert speech into an audio signal.

The effectiveness of the proposed model in generating high-quality synthesis across a wide range of speakers is likely due to the quality of the speaker encoder's learned representation. To investigate this further, we evaluated the impact of the speaker encoder training set on the quality of synthesis. We utilized the LibriSpeech training set, which contains 100 hours of speech data from 250 speakers who are not related to those in the clean subsets. In order to avoid overfitting, the speaker encoders trained on limited datasets (top two rows) used a more compact network architecture consisting of 256-dimensional LSTM cells with 64-dimensional projections and produced 64-dimensional speaker embeddings.

First, we assess the speaker encoder that was developed using the LibriSpeech Clean audio dataset, which have a comparable number of speakers. With the exception of being trained discriminatively, speaker encoder and synthesiser share a baseline with the speaker encoder that has not been fine-tuned.

The text transcripts and encoder embeddings are used to train the synthesizer to produce mel-spectrogram frames. Pre-processing the text transcripts to produce a series of a phoneme or grapheme embeddings that reflect the linguistic content of the text is a standard step in the training process. After pre-processing, a loss function is constructed to encourage the synthesiser to produce mel-spectrogram frames that are similar to the ground truth mel-spectrograms for the given text and embedding. The encoder embeddings are also supplied as input, and the model is trained using a large dataset of matched text and mel-spectrogram frames.

V. RESULTS AND ANALYSIS

The TTS system accepts an audio file in the WAV or FLAC formats and a corresponding text file as inputs, and it then converts the inputs into high-quality speech. The system creates an audio file based on the input text when the user submits the files by pressing the submit button. Although the procedure can just take a little while to finish, the finished audio file is produced in WAV format.

In order to assess the performance of our model, we utilized the Mean Opinion Score (MOS) metric, as described in the original paper. Although some studies have utilized crowd-sourced evaluation for their experiments, due to resource constraints, we were unable to employ this method. Our evaluation protocol instructed users to rate the Naturalness and Similarity of a synthesized audio compared to a reference speaker. The term "Naturalness" refers to how closely the synthesized audio resembles human speech, while "Similarity" pertains to how closely the voice of the synthesized sample matches that of the reference speech. The rating scale ranged from 1 (Least Similar/Natural) to 5 (Extremely Similar/Natural), in increments of 0.5. We assessed the audio samples of 10 speakers, of which 5 were included in the training data, while the other 5 were not.

Our evaluation yielded the following results:

- Voice Naturalness: 4.12±0.27
- Speech Similarity: 3.73±0.34

These results provide valuable insights into the performance of our model and will inform future developments in this area. We are created a web interface using flask for user interaction. Below figure shows the interface for giving inputs and playing the output audio.

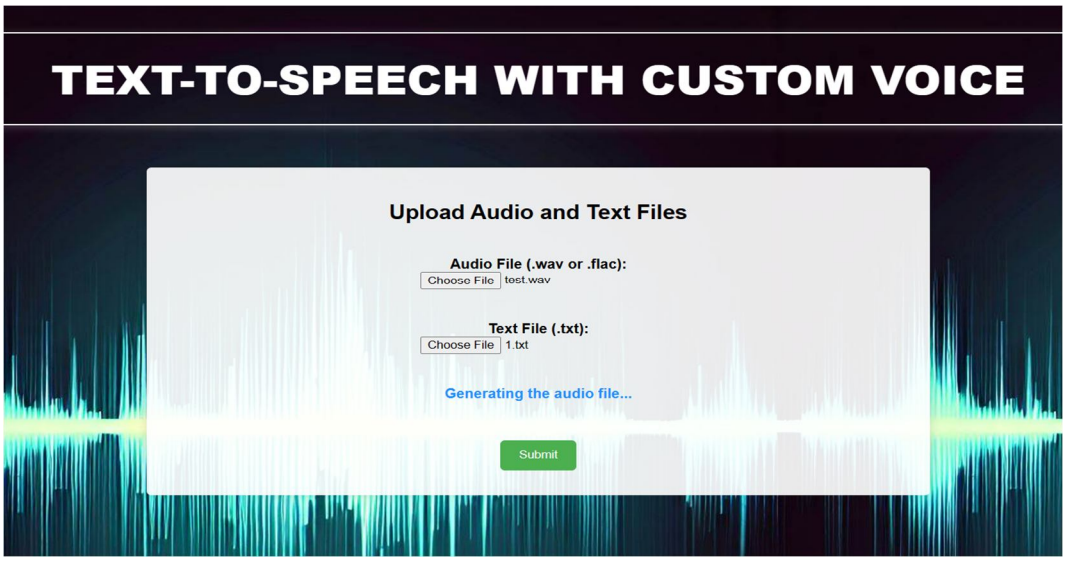


Fig.5.1 Input Page

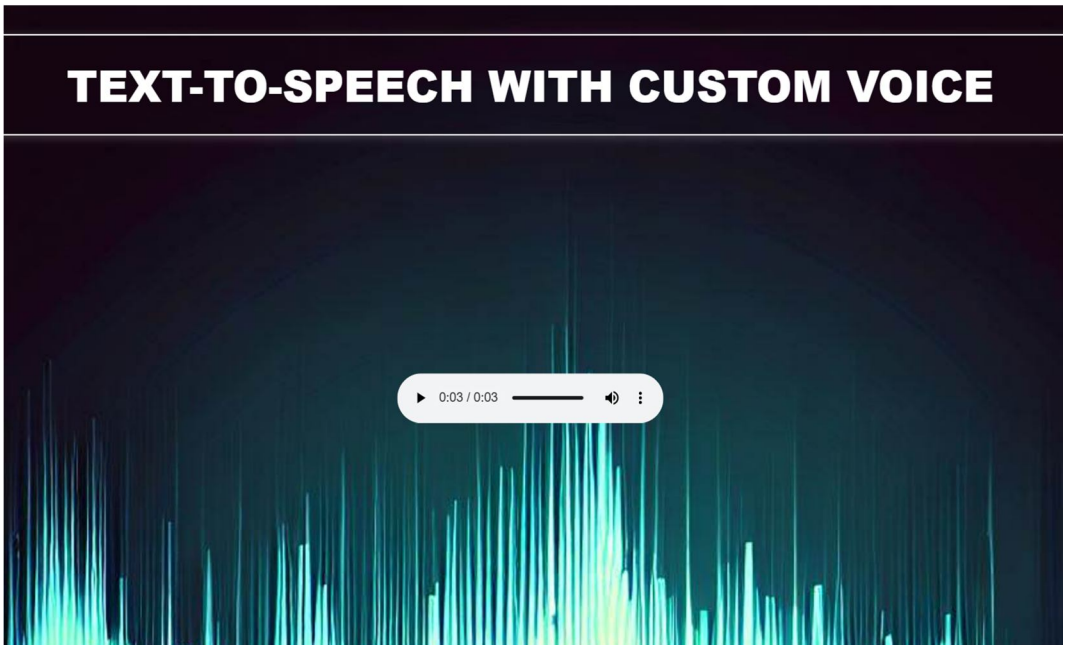


Fig.5.2 Result Page

VI. STRENGTH AND LIMITATION

A promising field of research is custom voice cloning due to its many advantages. It has the capacity to improve everyone's access to and personalization of digital content. When there is insufficient training data, having the capability to clone voices using a limited amount of data could prove to be useful. Additionally, the development of real-time voice cloning systems has a number of possible applications, including voice assistants, language learning, human-computer interactions, and assistive technology for those with visual impairments. Text-to-speech technology with custom voice has several advantages, but it also has some drawbacks. The quantity and quality of the training data, as well as the speaker's voice fluctuations, can all have an effect on how accurate voice cloning systems are. The use of voice cloning technology also raises moral questions about matters like privacy and consent, particularly if the technology is employed to create fake audio content for bad intentions. Additionally, the technology only supports English language at this time and has restricted language capabilities.

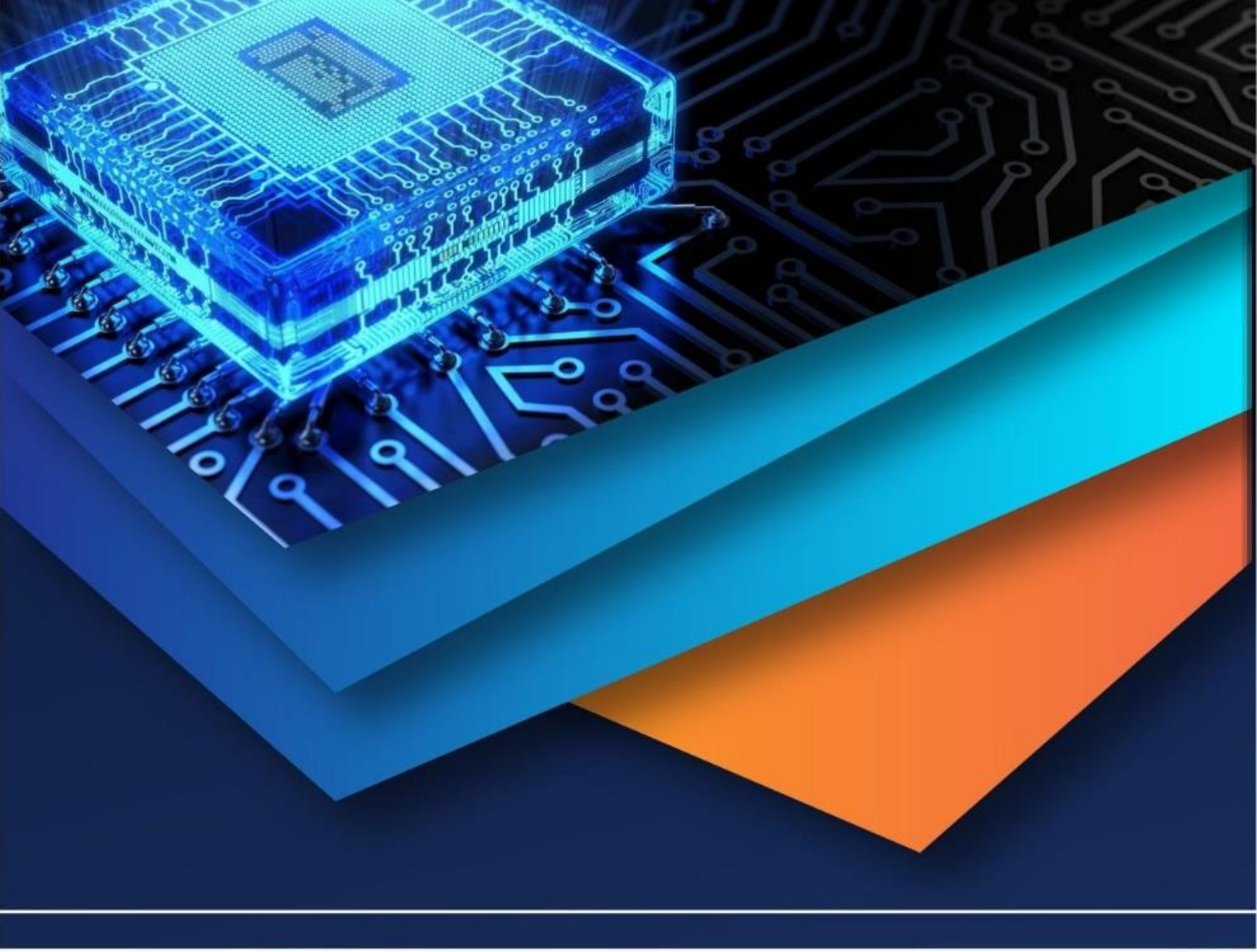
VII. CONCLUSION

In conclusion, new opportunities for producing distinctive and realistic-sounding voices have emerged as a result of the development of a text-to-speech system with custom voice cloning technology. The ability to produce high-quality customised voices that closely resemble the voice of a human speaker has become possible thanks to improvements in machine learning techniques and the accessibility of massive speech datasets. Applications for Text To Speech with custom voice system include more individualised audio content delivery, improved digital material accessibility, and enhanced human-computer interactions. The development of a real-time voice cloning system that can produce speech in real-time without any observable delays or errors, however, still faces a number of difficulties.

Future studies could concentrate on a number of areas to enhance the voice cloning system's functionality. The capacity of the system to produce speech that sounds more realistic is one of the main areas that needs improvement. The system's speech can currently sound robotic or mechanical, which makes it challenging for consumers to interact with the information. To solve this problem, researchers can look into adding more NLP approaches to the system in order to produce speech that is more expressive and flowing. The system could be improved further by adding support for more languages. The system now only supports English, which reduces the number of users it could accommodate. The technology could benefit a larger audience by enhancing its language skills to incorporate additional spoken languages, such as Indian languages. Additionally, future studies can concentrate on creating a system that is lighter and more effective and can be included into a variety of applications. The system's computational complexity might be reduced, and it could be made to work better in real-time, making it an important tool for improving the accessibility and customization of digital content.

REFERENCES

[1] "TACTRON: Towards end-to-end speech synthesis" - Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, Rif A. Saurous [August 2017].
[2] "Efficient Neural Audio Synthesis" - Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, Koray Kavukcuoglu [June 2018].
[3] "Transfer Learning from Speaker Verification to Multi-speaker Text-To-Speech Synthesis" - Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu [January 2019].
[4] "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System" - Weicheng Cai, Jinkun Chen, Ming Li [April 2018].
[5] "Char2Wav: End-to-end Speech Synthesis" - Jose M. R. Sotelo, Soroush Mehri, Kundan Kumar, J. F. Santos, Kyle Kastner, Aaron C. Courville, Yoshua Bengio [April 2018].
[6] "High-Fidelity and Low-Latency Universal Neural Vocoder based on Multiband WaveRNN with Data-Driven Linear Prediction for Discrete Waveform Modeling" - Patrick Lumban Tobing, Tomoki Toda [June 2021]
[7] "Combination of deep speaker embeddings for diarisation" - Guangzhi Sun, Chao Zhang, Philip C. Woodland [July 2018].
[8] "A comparative study of feature extraction techniques for speech emotion recognition" - Pratik K. Kurzekar, Ratnadeep R. Deshmukh, Vishal B. Waghmare, Pukhraj P. Shrishrimal [December 2014].
[9] "First step towards end-to-end parametric TTS synthesis: Generating spectral parameters with neural attention" - Wenfu Wang, Shuang Xu, Bo Xu [September 2016].
[10] "Speaker adaptation in DNN based speech synthesis using d-vectors" - Rama Doddipatla, Norbert Braunschweiler, Ranniery Maia [August 2017].



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY