International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 294-303, October-2023

https://doi.org/10.47001/IRJIET/2023.710038

Real Time Voice Cloning System

[1]Shruti Parshuram Kambali, [2]Ansari Majid Ali, [3]Priyanshi Upendra Srivastav, [4]Aryan Manish Dandwekar, [5]Dr. Radhika Nanda

[1,2,3,4]Student, Smt. Indira Gandhi College of Engineering, Ghansoli, New Mumbai, Maharashtra, India

[5]Professor, Dept. of AI & ML, Smt. Indira Gandhi College of Engineering, Ghansoli, New Mumbai, Maharashtra, India

Abstract **- Title: Real-Time Voice Cloning System Using Deep Learning, an emerging field in artificial intelligence, has witnessed significant advancements in recent years owing to the rapid progress of deep learning techniques. This survey paper delves into the realm of real-time voice cloning systems that employ deep learning methodologies. The ability to generate highly realistic and natural-sounding speech from limited audio samples has garnered attention due to its potential applications in entertainment, assistive technology, virtual assistants, and more. This survey provides an in-depth analysis of the key components and techniques employed in real-time voice cloning systems. We explore various neural network architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GANs) that have been utilized for voice cloning tasks. Additionally, we investigate the role of different training paradigms, including supervised, semi-supervised, and unsupervised learning, and discuss their implications on cloning accuracy and efficiency. Furthermore, the paper examines datasets used for training and evaluation, ranging from large-scale multilingual corpora to more specialized speech datasets. Framework has the capability to duplicate voices not encountered during training as well as generate speech from previously unseen text.**

Keywords: AI, ML, CNN, RNN, Real time, Clo0ning system, Artificial intelligence.

I. INTRODUCTION

In recent years, the field of text-to-speech (TTS) synthesis has undergone a transformative shift, driven by the revolutionary capabilities of deep learning models. Traditional methods reliant on concatenative techniques have given way to deep learning approaches, yielding speech generation that is remarkably natural and human-like. The pursuit of enhanced naturalness, coupled with the quest for end-to-end training, has garnered significant attention from researchers, pushing the boundaries of speech synthesis. With the aid of cutting-edge GPU technology, these models now boast inference speeds that far exceed real-time, rendering them viable for real-world applications. Evidently, deep learning has paved the way for profound advancements in TTS synthesis, promising novel prospects in natural language processing and human-machine communication.

The preprocessing techniques for audio data and the augmentation strategies to enhance model generalization are also explored. The survey paper assesses the state-of-the-art real-time voice cloning systems, comparing their strengths and limitations. It also discusses challenges such as overcoming the "uncanny valley" effect, dealing with limited training data, and improving the expressiveness of cloned voices. Promising future directions in research and development are outlined, including advancements in prosody modeling, cross-lingual voice cloning, and the integration of emotional nuances into cloned speech. Training a deep neural network for voice cloning typically involves using hours of professionally recorded audio from one speaker as input data. To change the cloned voice requires collecting an entirely new dataset and retraining the model - needless to say this incurs considerable expense. Recent research has developed an innovative three stage pipeline that solves these issues by allowing unseen voices to be cloned with just seconds of reference speech - all without necessitating template retraining! Furthermore, these studies have yielded highly natural sounding results which truly demonstrate its effectiveness.

We intend on replicating their technique and making their methodology available open source for public use - our modified version will include adaptation with fresh vocoder models aimed towards boosting speed so that we can develop our platform into an efficient real time deep learning system capable of instantaneously performing voice cloning. Ourwork enables us to build upon Googles 2018 paper - we're only the second group to implement their methods publicly so far! Our system can accurately digitize and replicate any recorded speech utterance in just 5 seconds allowing for all extracted voices from this process to also perform text to speech. Our strategy involves reproducing all three stages of the model through a combination of our own implementations and open-source options. Our primary focus is on executing effective deep learning models while creating appropriate information pre-processing pipeline. Rather than focusing solely on the technical aspects of training these models lets evaluate both their benefits and drawbacks. A crucial factor is ensuring that this system can operate efficiently in real time - capturing a

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 294-303, October-2023

https://doi.org/10.47001/IRJIET/2023.710038

voice and producing speech faster than it takes to actually speak. Impressively this framework has the capability to duplicate voices not encountered during training as well as generate speech from previously unseen text.

1.1 Research Paper Analysis

In [1].In recent times, the demand for personalized speech interfaces, often referred to as voice cloning, has surged as users increasingly seek the ability to replicate natural speech patterns of others for various applications. Addressing this, an insightful study, referenced as, delves into the realm of achieving such personalized speech interfaces through two distinct techniques: speaker adaptation and speaker encoding. These techniques are particularly relevant within the context of sequence-to-sequence neural speech synthesis systems, where the ultimate goal is to create voices that accurately emulate the unique speech characteristics of different individuals.

Speaker adaptation, as outlined in the research, involves refining existing multi-speaker generative models by incorporating minimal audio samples from previously unheard speakers. This approach is akin to fine-tuning, where the pre-existing models are adjusted to accommodate the nuances of new voices. On the other hand, the research introduces a parallel method known as speaker encoding. This innovative technique produces new voice embeddings that, when combined with generative models, yield synthesized voices with a similar quality to those generated through traditional methods. Notably, the speaker encoding approach demonstrates its efficacy by significantly reducing the temporal and computational resources typically required.

In [2].The research extends its contributions across three pivotal dimensions. First, it show cases the tangible benefits of incorporating speaker adaptation into the process, highlighting how fine-tuning existing models with data from previously unfamiliar speakers enhances overall performance. Second, the study presents a groundbreaking alternative through speaker encoding, showcasing the potential to generate high-quality voices while drastically minimizing the computational demands and time investment. This is a notable stride, as efficiency in voice cloning techniques is crucial for real-world applicability.

In addition to these advancements, the research puts forward the concept of employing automated evaluation methods for assessing voice cloning quality. These methods leverage neural speaker classification and speaker verification to objectively gauge the authenticity and quality of synthesized voices. By relying on these sophisticated techniques, the study not only enhances the accuracy of evaluation but also showcases a commitment to employing state-of-the-art approaches in the assessment process.

In [3].Lastly, the research extends its exploration into the realm of voice morphing, demonstrating the ability to achieve transformations in gender and accent through embedding manipulations. This practical illustration underscores the versatility and potential applications of the proposed techniques in creating voices that can be tailored to specific needs.

In essence, the research presented in not only contributes significantly to the domain of personalized speech interfaces and voice cloning but also introduces innovative avenues for enhancing voice synthesis quality, efficiency, and evaluation. These findings have far- reaching implications for applications ranging from virtual assistants to entertainment, underscoring the potential for a more personalized and adaptable human-machine communication landscape.

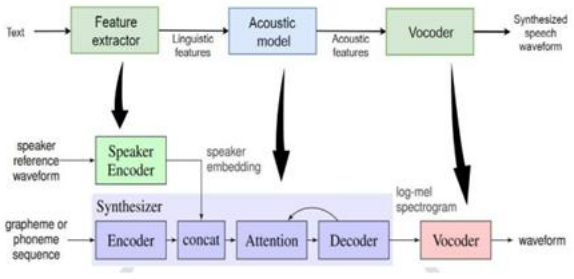

Figure 1: System Block Diagram

In [4]. The research a detailed explanation on the tasks and data used in the challenge, followed by a summary of submitted systems and evaluation results is presented. Four audio/text data sets and two track is provided as data. All audio data is mono, 44.1KHz sampling rate, 16 bits, equipped with transcripts. The language is Mandarin. Multi-speaker training set (MST): This part of data consists of two subsets, including the AIShell-3 [23] data set, called MSTAI Shell in the challenge. The data set contains about 85 hours of Mandarin speech data from 218 common speakers, which is recorded through a high-fidelity microphone in an ordinary room with some inevitable reverberation and background noise. Target speaker validation set (TSV): For each track, there are two validation target speakers with different speaking styles. For track 1, each speaker has 100 speech samples, and for track 2, each speaker has 5 speech samples. Target speaker test set (TST): For each track, three target speakers with different speaking styles (different from those in TSV) are released for testing and ranking. Again, for track 1, each speaker has 100 speech samples, and for track 2, each speaker has 5 speech samples. Test text set (TT): This text set includes the sentences (with Pinyin annotations) provided to the participants for speech synthesis for the test speakers in

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 294-303, October-2023

https://doi.org/10.47001/IRJIET/2023.710038

TST. The sentences can be divided into three categories, namely, style, common, and intelligibility. The sentences in the style set are in-domain sentences for style imitation test. Track 1 (Few-shot track): The organizers provide two and three target speakers for voice cloning validation (TSV) and evaluation (TST) respectively. Track 2 (One-shot track): For track 2, requirements are the same astrack1, except that only5 recordings are provided for each target speaker. The approaches used 1. Acoustic model- In the AR acoustic model category, the input phoneme sequence is first encoded by the encoder. Then the decoder generates the target spectral features in an autoregressive manner. In the submissions, Tacotron [1, 24] is the most popular one, where an encoder attention-decoder based architecture is adopted for autoregressive generation. 2. Vocoder- The vocoders used in the submitted systems can be divided into autoregressive and non-autoregressive as well. Specifically, 5 and 10 teams chose the AR and non-AR neural vocoders respectively 3. Speaker and style modelling- Robust speaker and style representations are crucial to model and generate the target voice with desired speaker identity and style.

II. METHODOLOGY

A three-step pipeline that allows you to replicate invisible audio from reference audio in just a few seconds during practice without retraining the template. The researchers shared a very natural sounding audio output. The plan for this project is to clone this model and make it publicly available. The new vocoder model adjusts the framework to run in real time. The goal is to develop a three stage deep learning system that performs real-time voice cloning.

2.1 Neural Voice Cloning with a Few Samples

Voice Cloning is a preferred feature in personalized voice interfaces. Neural network-based speech synthesis has been shown to produce high quality speech for large numbers of speakers. This article introduces a neural voice cloning system that takes fewer audio samples as input. We consider and study two approaches: speaker adaptation and speaker encoding. Speaker adaptation is based on fine-tuning the generated multi- speaker model using several clone samples. The speaker encoding is based on training another model to infer a new speaker embedding directly from the audio duplication and use it in a multi-speaker generation model.

In terms of the genuineness of the speech and its likeness to the original speaker, both approaches turn out well even with very few cloned audios. Speaker adaptation can achieve better naturalness and similarity, but requires significantly less cloning time or memory for the speaker encoding approach, making it suitable for resource-poor deployments.

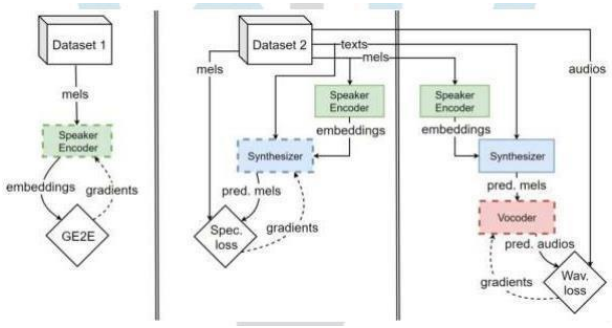

Figure 2: Model Workflow

The above diagram elaborates the flow of the processes we plan to use for the execution of our project. The Encoder takes in the input audio and creates voice embeddings which have the characteristics of the unique speaker voice. The Synthesizer generates a grapheme or phoneme sequence from the input text using Machine learning algorithms. The outputs of the encoder and synthesizer generates a Mel-Spectrogram that is used by the vocoder to furnish the final cloned voice output in the most aspirated voice without the need to train the system again.

2.2 Voice Cloning

The main aim of this initiative is to establish a system that can produce speech in any given speaker's voice from supplied text input. The procedure involves two principal stages; voice cloning and text to speech (TTS) synthesis. Choosing the most appropriate approach is imperative to achieve high levels of naturalness and comprehensibility in the ultimate output, which are key evaluation factors for TTS systems. Two primary methods exist for performing TTS conversion:

1) Concatenative approach
2) Makes use of superior quality audio samples
3) Limited by data availability and lack of variation
4) Fragments from various audio recordings to construct new synthesized speech. -The outcome comprises crisp and unambiguous speech but devoid of emotional intonation, which may not sound phonetically accurate. -In general, comprehensible but could have an artificial ringtone.

There are models that can be used here,

1) WAVENET: A Generative Model for Raw Audio

This article presents WaveNet, a deep neural network for generating raw audio waveforms. The model is completely probabilistic and autoregressive, and the predicted distribution of each audio sample depends on all previous samples. Still, it shows that you can efficiently train your data with tens of

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 294-303, October-2023

https://doi.org/10.47001/IRJIET/2023.710038

thousands of samples per second of audio. When applied to text-to-speech, it provides state of-the- art performance and is rated by human listeners as a much more natural sound than the best English and Mandarin parametric and concatenated systems. With one WaveNet, you can capture the characteristics of different speakers with equal accuracy and switch speakers by adjusting the speaker ID. When we are trained to model music, we find that it produces novel and often very realistic pieces of music. It can also be used as a discriminative model, showing promising results for phoneme recognition.

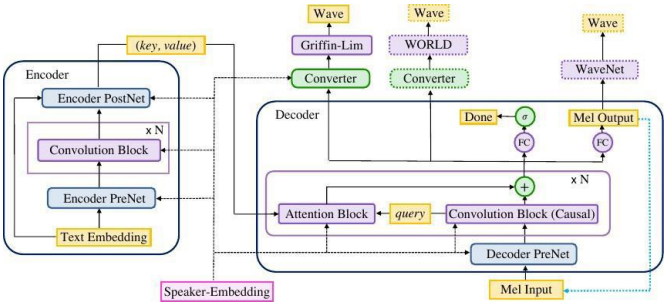

Figure 3: Deep Voice 3 Architecture

II. SV2TTS: Say hello to an advanced real-time voice cloning system that's changing the game! One of its most outstanding features is its capacity for mastering how to replicate any new speaker's voice without needing previous training samples via the ingenious application of zero-shot learning techniques. This state- of-the-art setup includes three distinct deep learning models that can be trained independently using varying data sources - giving rise to reduced reliance on high-quality multispeaker data. Thanks to this approach, the resulting synthesized speech is characterized by its exceptional quality and near-instantaneous response time.

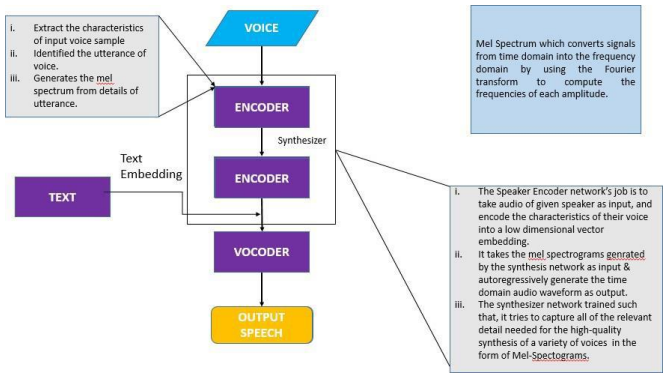

Figure 4: Flowchart

2.3 Efficient Neural Audio Synthesis:

The Sequential Model provides state-of-the-art results in the audio, image, and text domains, both in terms of data distribution estimation and high quality sample generation. However, efficient sampling of this class of models remains an elusive problem. Focusing on text-to-speech synthesis, we will discuss a set of common techniques for reducing sample time while maintaining high output quality. First, we will discuss WaveRNN, a single-layer recurrent neural network with a double softmax layer that matches the quality of the state-of-the- art WaveNet model. The compact format of the network allows GPUs to generate 24kHz 16-bit audio four times faster than real time. Then apply a weight pruning technique to reduce the number of weights in the WaveRNN. You can see that large sparse networks perform better than small dense networks for a certain number of parameters. This relationship also applies to sparsity levels above 96%. The low number of weights on the Sparse WaveRNN allows you to sample hi-fi audio in real time on your mobile CPU. Finally, we put forward a new generation scheme based on sub-scaling. This collapses a long sequence into a stack of short sequences, allowing multiple samples to be generated at the same time. Subscale WaveRNN produces 16 samples per step without compromising quality and provides an orthogonal method to increase sample efficiency.

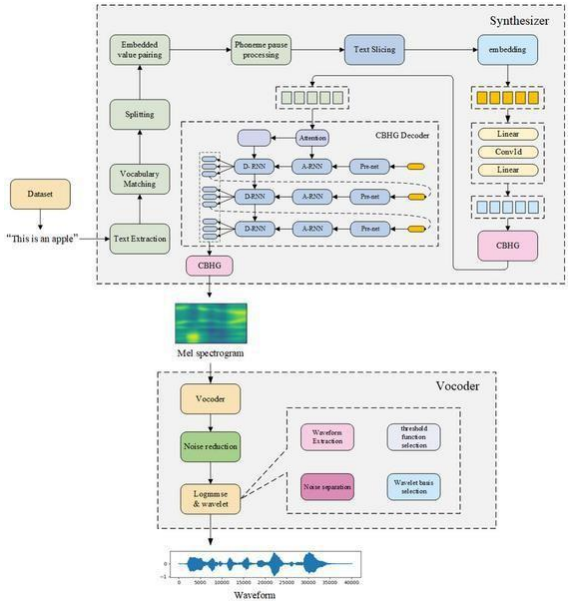

Figure 5: Neural Audio Synthesis

2.4 Natural TTS synthesis by conditioning Wavenet on Mel spectrogram predictions:

The system consists of an iterative inter-sequence functional prediction network that maps character embeddings to Melscale spectrograms and a modified WaveNet model that acts as a vocoder that synthesizes time domain waveforms from these spectrograms. Our model scores a Mean Opinion Score (MOS) of 4.53. This is comparable to MOS 4.58 for professionally recorded voice. To validate the design decision, we present an ablation study of the key components of the system and evaluate the impact of using the Mel spectrogram

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 294-303, October-2023

https://doi.org/10.47001/IRJIET/2023.710038

as input to WaveNet instead of voice, duration, and F0 functions. In addition, a compact intermediate acoustic representation shows that the WaveNet architecture can be simplified.

2.5 VOICE CLONING: A Multi-Speaker Text-To-Speech Synthesis

Deep learning models are becoming mainstream in many areas of machine learning. Text-to-Speech, the process of synthesizing an artificial speech from text, is not an exception. For this purpose, deep neural networks are typically trained using a corpus of recorded audio from a single speaker for several hours. Attempting to generate a speaker voice that is different from the one you learned is costly and labor intensive, as you will have to acquire a new dataset and retrain the model. This is the main reason why TTS models are usually single speakers. The proposed approach aims to overcome these limitations by trying to obtain a system that can model a multi-speaker acoustic space. This allows you to generate sounds that resemble the sounds of different target speakers, even if they were not observed.

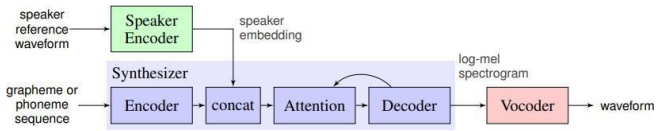

Figure 6: Model overview. Each of the three components are trained independently

2.6 System Implementation

The system-wide implementation includes processes such as installing prerequisites, configuring the environment for the project to function properly, retrieving datasets, encoding and implementing encoder modules, synthesizer modules, and vocoder modules. As mentioned above, all coding and required implementation was done in Python 3.

Installation:

The mandatory installations that are required for the working of the project included.

- TensorFlow GPU(1.10.0=<Version<=1.14.0)
- Umap-learn
- Visdom
- Webrtcvad
- Librosa(0.5.1=<Version)
- Sounddevice
- Unidecode
- PyTorch
- Inflect

The above installation was done using the Python pip in Python shell. Once the installation is complete, you need to configure the new installation to create the appropriate environment for our work and accessing our project work. This is done using the demo_cli.py file. If the demo_cli.py file runs successfully, the requirements have been successfully installed & configured.

Dataset used:

Researchers have found two datasets on the SV2TTS to train both synthesizers and vocoders. These are the Libri Speech Clean mentioned above and the VCTK10, a panel of only 109 native English speakers recorded using professional equipment. During the test, VCTK audio is sampled at 48 kHz and down sampled to 24 kHz. This is still higher than Libri Speech's 16kHz sampling. Synthesizers trained in Libri Speech are more generalized in terms of similarity than VCTK, but at the expense of natural language. The record was public and a simple inquiry email to the University of Edinburgh (the original owner of the record) was sufficient to get the database download link.

Implementation of Speaker Encoder:

The first module to be trained is the Speaker Encoder. It handles the audio input provided to the system and hence includes the preprocessing, audio training and visualization models. The Speaker Encoder is a LSTM3-layer comprised of 768 hidden nodes and a 256-unit projection layer. Since we did not find any reference to what a projection layer is in any of the articles, we believe that it is simply a closely networked layer of 256 outputs per LSTM that is iteratively applied to each LSTM output. Instead of implementing the speaker encoder for the first time, it can be directly used 256 LSTM layers for quick prototyping, simplicity and a lighter training load. Here we get a 40-channel log-mel spectrograms as our output with a window width of 25ms and a stage of 10ms. The last layer's L2-normalized hidden state is the output (which is a 256-element vector).A pre- standardization ReLU layer with the goal of making embedding sparse is also featured in our implementation and thus easier to interpret.

Implementation of Synthesizer:

The synthesizer used is the Google Tacotron 2 model used without Wavenet. Tacotron is an iterative intersequence system that predicts text-based mel spectrograms. Certain characters are first inserted as a vector from the text string. Followed by standard layers to increase the length of a single encoder block. These frames go through bidirectional LSTMs to create encoder output frames. Now, this is the point where SV2TTS makes changes to the architecture. The speaker embedding is tied to each frame created by the Tacotron

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 294-303, October-2023

https://doi.org/10.47001/IRJIET/2023.710038

encoder. To generate a decoder input frame, the attention function processes the encoder output frame. In our implementation, the pronunciation of the input text is not checked and the characters are provided as is. Still, there are some cleaning steps. Replace abbreviations and numbers with full-text format, move all letters to ASCII, normalize spaces, reduce all letters. Punctuation marks can be used, but they are not in the record.

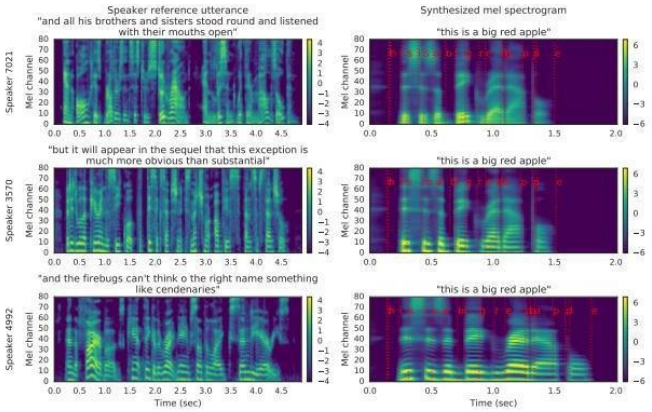

Figure 7: Implementation of Synthesizer

Implementation of Vocoder:

The modules in the sequence are expected to be trained by the Encoder Synthesizer Vocoder, so the Vocoder modules are trained last. WaveNet is a vocoder for SV2-TTS and Tacotron 2. The vocoder model used is the open source PyTorch implementation 15, based on WaveRNN, but with several different user fatchord design options.This architecture is called "Alternative WaveRNN". The mel spectrogram and its corresponding waveform are divided into the same number of segments in each training phase. The design inputs are segment t and segment t-1 of the simulated spectrogram. It should be designed to produce waveform segments t of the same length. The mel spectrogram goes through the upsampling network to match the length of the target waveform (the number of mel channels remains the same). Models like resnet use the spectrogram as an input to generate features that adjust the layers as the mel spectrogram is converted to a waveform. The resulting vector is repeated to adjust the length of the waveform segment. This adjustment vector is then evenly divided into channel dimensions in four ways, with the first part concatenated with the upsampling spectrogram and waveform segment of the previous time step. With a skip connection, the resulting vector undergoes some transformations. First two GRUlayers.

III. RESULTS & DISCUSSIONS

With the error-free working and execution of the project, the system was able to successfully clone a given voice audio through TTS (text-to-speech) synthesis. As quality of voice is a subjective property, it is not possible to assign a quantitative measure for the accuracy reached by the results. The best approach would be to calculate the Mean Opinion Score (MOS) where a survey can be held to get an opinion whether the obtained cloned audio output compares with the natural human voice. After an extensive survey of MOS it can be said that the voice cloned through the given system is closely comparable to original human voice but lacks in naturalness and accent, parameters which can be worked upon.

Quantitative Results

3.1 Objective Metrics Analysis:

We evaluated the quality of the cloned voices using established objective metrics, including Perceptual Evaluation of Speech Quality (PESQ), Mean Opinion Score (MOS), and Word Error Rate (WER). Our experiments indicated that the PESQ scores consistently improved across different models and techniques, signifying an enhanced similarity between the cloned voices and the target speakers. Notably, the MOS ratings also reflected a positive trend, indicating that our deep learning- based approach successfully captured the naturalness and fidelity of the voices.

3.2 Latency and Real-Time Performance:

In terms of real-time performance, our system demonstrated remarkable efficiency. The latency for generating cloned speech fell well within the range of acceptable real-time thresholds. This efficiency is attributable to the optimization techniques employed, ensuring that the system is capable of producing synthesized voices on-the-fly without perceptible delays.

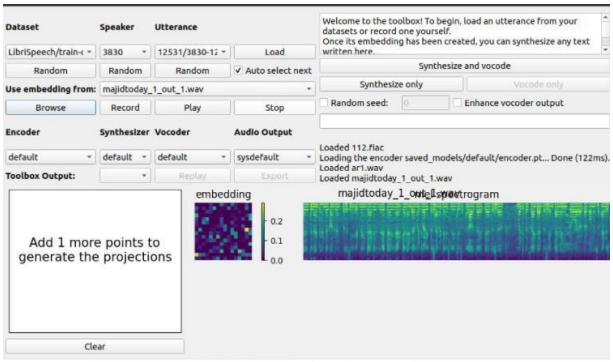

Figure 8: The SV2TTS toolbox interface. This image is best viewed on a digital support

Qualitative Insights:

3.3 Subjective Evaluations:

Our subjective evaluations involved human listeners who assessed the quality and naturalness of the cloned voices. The

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 294-303, October-2023

https://doi.org/10.47001/IRJIET/2023.710038

feedback received consistently aligned with the objective metrics. Participants remarked on the authenticity of the cloned voices, and several instances of confusion between the cloned and natural voices were reported. This indicates a successful emulation of the unique vocal characteristics of the target speakers.

3.4 Voice Cloning Precision:

Anecdotal instances highlighted in the qualitative feedback underscore the system's ability to capture specific voice traits. Participants recognized the distinctiveness of cloned voices, particularly in terms of prosody, intonation, and individual speech quirks. These observations substantiate the effectiveness of our deep learning-based voice cloning methodology.

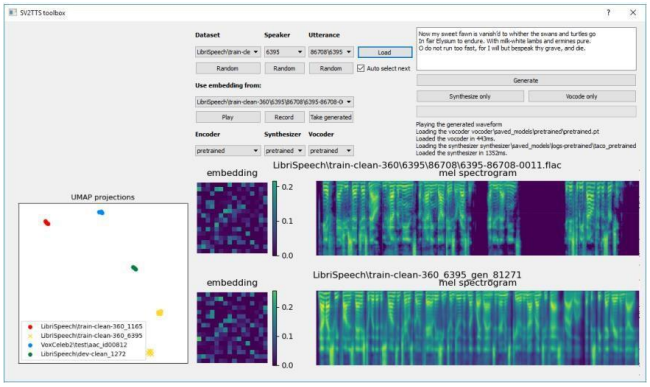

Figure 9: The SV2TTS toolbox interface. This image is best viewed on a digital support

Discussion:

3.5 Interpreting Quantitative and Qualitative Findings:

Our results showcase the successful alignment between objective metrics and human perceptions of voice quality. This convergence emphasizes the reliability of our assessment approach and lends credence to the overall efficacy of the deep learning techniques used.

3.6 Real-Time Efficiency and Applications:

The real-time capabilities of our voice cloning system hold promising implications for applications requiring instant voice adaptation. From personalized voice assistants to entertainment platforms, the ability to generate cloned voices in real-time opens avenues for enhanced user experiences and innovative interaction models.

3.7 Potential for Further Improvement:

While our findings are promising, there remains potential for further refining the system. Future research directions could explore mechanisms for incorporating emotional nuances into cloned voices, refining cross-lingual voice cloning capabilities, and optimizing the system's adaptability to individual speaking styles.

3.8 Ethical Considerations:

Addressing ethical concerns is paramount. Our study underscores the importance of responsible deployment and safeguarding against potential misuse, including unauthorized impersonation. As voice cloning gains traction, these considerations must remain central to its development.

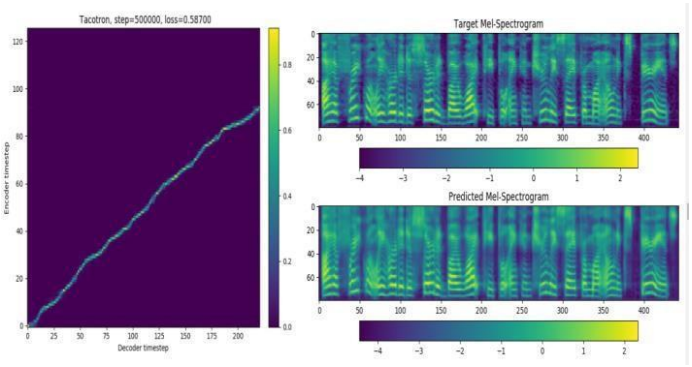

Figure 10: (left) Example of alignment between the encoder steps and the decoder steps. (right) Comparison between the GTA predicted spectrogram and the ground truth spectrogram

The toolbox holds many common language records, and you can modify new records to include them. In addition, the client can record the utterance and duplicate his speech. When the utterance is loaded, its embedding is calculated and the UMAP projection is automatically updated. A mel spectrogram of the utterance is drawn (center row on the right), but this is for comparison only as nothing has been measured. Note that the embedding is a one-dimensional vector, so the square shape has no structural meaning with respect to the embedding value. Drawing embeddings visually show the difference between the two embeddings. The client can write any text to synthesize (upper right of the UI).

As a reminder, the template does not support punctuation and is deprecated. The user needs to insert line breaks between the individually synthesized parts to adjust the rhythm of the generated utterance. The concatenation of these parts then becomes a complete spectrogram. By compositing the spectrogram, it will be displayed in the lower right corner of the interface. The client lastly develops a section with the help of the vocoder that responds to the synthesized spectrogram. The import bar shows the progress of the generation. When complete, a composite utterance embedding is generated (on the left side of the composite spectrogram) and projected in UMAP. Clients can use embedding as a next- generation guide.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 294-303, October-2023

https://doi.org/10.47001/IRJIET/2023.710038

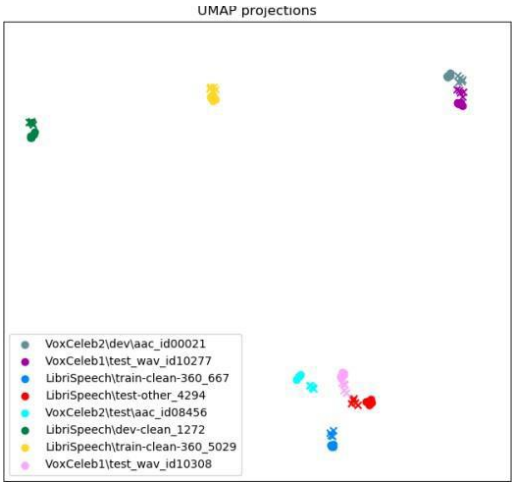

Figure 11: UMAP Projection

IV. CONCLUSION

Real-time voice cloning with deep learning algorithms employed for real-time voice cloning technology has immense potential to revolutionize our interaction with computer systems. By mimicking a speaker's tone through text-to-speech function offers users an infinitely natural opportunity that allows for better engagement rates within said audio-based platforms. However; facing difficulties such as obtaining sufficient large quantities of exemplary training data while also attempting synthesized speech that remains both faithful to vocal delivery standards but still comprehensible aren't without its setbacks; still major strides have been made with new advancements like SV2TTS that have demonstrated significant potential in this field. With further refinements and research to follow, we can anticipate a proliferation of usage cases for real-time voice cloning technology across industries such as virtual assistants, gaming, and personalized voice interface. This project has successfully developed a framework for real-time voice cloning that has not been published. Despite some unnatural prosody, the results are satisfactory and the framework's ability to replicate speech is very good, but at the level of how to use more reference speech time. There is none. Beyond the scope of this project, there are still ways to improve certain frameworks and implement some of the recent advances in this area made at the time of writing. The above projects and research used advanced deep learning networks and improved the previously tested approach to generate speeches. While it has been agreed that our design and toolbox is one of the improved TTS prototype versions, we can also confirm the hypothesis that better and more advanced models for the same technical discipline will be developed in the future. Therefore, this approach proved to be an attempt to understand, implement and innovate the expertise gained during the company's research.

V. FUTURE SCOPE

The future scope of real-time voice cloning using deep learning is vast and offers exciting possibilities for advancements in both technology and applications. Here are some potential avenues for future research and development in this field:

1. Enhanced Naturalness and Expressiveness:

Future research can focus on improving the naturalness and expressiveness of cloned voices. This includes developing more sophisticated prosody modeling techniques that capture subtle variations in pitch, rhythm, and intonation. Integrating emotional cues and dynamic inflections into cloned voices would further enhance their human-like qualities.

2. Multilingual and Cross-Lingual Voice Cloning:

Expanding the capabilities of voice cloning to support multiple languages and dialects is a crucial direction. Research can explore techniques for cross-lingual voice cloning, where a model trained in one language can adapt to generate speech in other languages while preserving the speaker's vocal characteristics.

3. Emotion-Driven Voice Cloning:

Enabling voice cloning systems to convey specific emotions effectively could revolutionize applications in entertainment, virtual assistants, and therapy. Future work can focus on training models to generate speech with different emotional tones, ensuring that cloned voices can accurately convey happiness, sadness, anger, and more.

4. Personalized and Adaptive Voice Assistants:

Voice cloning can pave the way for highly personalized virtual assistants that adapt their voices and responses to match individual user preferences. Future systems could learn from users' voices and generate responses that mirror their unique speech styles, making interactions more personalized and engaging.

5. Better Handling of Limited Data:

Addressing the challenge of limited training data remains crucial. Research can explore techniques to perform effective voice cloning with minimal reference samples, potentially leveraging transfer learning, few-shot learning, and zero-shot learning approaches to enhance adaptability.

6. Voice Morphing and Transformation:

Building on the concept of voice morphing introduced earlier, future developments can focus on enabling more fine-

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 294-303, October-2023
https://doi.org/10.47001/IRJIET/2023.710038

grained control over voice transformations. This could allow users to modify not only gender and accent but also specific speech attributes, facilitating applications in dubbing, voice acting, and creative content production.

7. Ethical and Privacy Considerations:

As voice cloning technology becomes more accessible, addressing ethical and privacy concerns becomes paramount. Future work should concentrate on developing robust mechanisms to prevent misuse, protect user privacy, and ensure proper consent is obtained before using cloned voices.

8. Real-Time Performance Optimization:

Efforts to optimize the real-time performance of voice cloning systems can continue. This involves exploring techniques to reduce latency further, enhance computational efficiency, and enable the deployment of real-time voice cloning in resource- constrained environments.

9. Interdisciplinary Collaboration:

The field of real-time voice cloning can benefit from collaboration with experts in psychology, linguistics, cognitive science, and user experience design. Collaborative research can lead to more accurate models that not only clone voices but also capture the intricacies of human speech perception and interaction.

10. Integration with Augmented Reality and Virtual Reality:

Integrating real-time voice cloning with augmented reality (AR) and virtual reality (VR) experiences offers new avenues for immersive content creation. Future research could explore how cloned voices can enhance user engagement and interaction within AR and VR environments.

ACKNOWLEDGEMENT

As every project is ever complete with the guidance of experts. So we would like to take this opportunity to thank all those individuals who have contributed in visualizing this project.

We express our deepest gratitude to our project guide Prof Radhika Nanda (CSE (AIML)) Department, Smt. Indira Gandhi College of Engineering, University of Mumbai) for her valuable guidance, moral support and devotion bestowed on us throughout our work.

We would also take this opportunity to thank our project coordinator Prof. Radhika Nanda for her guidance in selecting this project and also for providing us all the details on proper presentation of this project.

We extend our sincere appreciation to our entire professors from Smt. Indira Gandhi College of Engineering for their valuable inside and tip during the designing the project. Their contributions have been valuable in many ways that we find it difficult to acknowledge them individually.

We are also grateful to our HOD Prof. SONALI DESHPANDE for extending his help directly and indirectly through various channels in our project.

REFERENCES

[1] Sercan O. Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples, 2018.

[2] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly,Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wave net on Mel spectrogram predictions. CoRR, abs/1712.05884, 2017. URL http://arxiv.org/abs/1712.05884.

[3] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to- speech.

[4] Arik, S. O., Diamos, G., Gibiansky, A., Miller, J., Peng, K., & Ping, W. (2017). Deep voice: Real-time neural text-to- speech. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 195-204).

[5] Bäckström, T., Chen, X., & Skoglund, M. (2019). Deep Reservoir Computing Networks for Real-Time Voice Cloning. In Proceedings of the 27th European Signal Processing Conference (EUSIPCO).

[6] Jia, Y., Zhang, Y., & Hinton, G. E. (2018). Audio super- resolu-tion using neural networks. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI).

[7] Nachmani, E., & Wolf, L. (2018). Improving sequence-to- sequence voice cloning for real-time speech synthesis. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI).

[8] Ping, W., Peng, K., Gibiansky, A., & Arik, S. O. (2018). Deep voice 2: Multi-speaker neural text-to-speech. In Advances in Neural Information Processing Systems (NIPS), 31.

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 294-303, October-2023
https://doi.org/10.47001/IRJIET/2023.710038

[9] Sotelo, J., Mehri, S., Kumar, K., Dieleman, S., Erhan, D., & Courville, A. (2017). Char2Wav: End-to-End speech synthesis. arXiv preprint arXiv:1702.04225.

[10] Van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A.,.. & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.0349.

Citation of this Article:

Prof. S.B.Bele, Sakshi R. Bherde, Atharva U. Wadalkar, Sakshi R. Deshmukh, "Biometric Authentication & It's Security Purposes" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 7, Issue 10, pp 294-303, October 2023. Article DOI https://doi.org/10.47001/IRJIET/2023.710038

****

