INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 05, May 2023, pp : 1130-1133

e-ISSN : 2583-1062

Impact Factor : 5.725

THE REAL TIME VOICE CLONINIG

M Mohanapriya[1], Daram Guru Prasad[2], Shashanka K S[3], Polakala Somasekhar[4], Akshay V Balihallimath[5]

[1]Assistant Professor, CSE, Sambhram Institute of Technology, Bengaluru, Karnataka, India

[2,3,4,5]Student, CSE, Sambhram Institute of Technology, Bengaluru, Karnataka, India

ABSTRACT

A recent study introduces a three-level pipeline for real-time voice cloning, capable of generating natural sounding speech from a few seconds of reference speech without retraining. The framework is based on Google's 2018 paper and aims to be open-sourced. The system can capture and generate audio in real-time and create a practical representation of the voice spoken in a virtual layout. The next step is to train the models with large datasets of tens of hours of audio and observe their strengths and weaknesses. The system can also generate audio from unseen text content, making it a promising tool for text-to-speech applications.

Key Words: The Real Time Voice Cloning, Text to Speech, Voice Conversation, Neural Networks, Speaker Encoding.

1. INTRODUCTION

Real-time voice cloning is the process of generating a synthetic voice that sounds like a particular person's voice in real-time. It is accomplished by training machine learning algorithms on audio recordings of the target person's voice. These algorithms use deep neural networks to learn the unique patterns and nuances of the person's voice, including their tone, pitch, accent, and speaking style. Once the algorithm has been trained, it can generate synthetic voice samples in real-time that closely mimic the target person's voice.

This technology has a variety of potential applications, such as in virtual assistants, text-to-speech synthesis, and voice acting. However, it also raises concerns around the potential misuse of synthetic voices for impersonation, fraud, and other unethical purposes. As such, it is important to consider ethical and legal implications as this technology continues to develop and become more widely available.

Overall, real-time voice cloning is an exciting development in the field of AI, with both positive and negative implications. It has the potential to revolutionize various industries and improve the lives of people with speech impairments, but it also requires responsible use and ethical considerations.

2. PROBLEM STATEMENT

The project should address the following challenges:

Voice Acquisition: Create a mechanism to collect a significant amount of high-quality audio data from the target speaker touse as training material for the voice cloning model.

Voice Representation: Develop a suitable representation for capturing the unique characteristics of a person's voice, including pitch, timbre, and intonation. This representation should be capable of capturing both the linguistic and emotional aspects of speech.

Deep Learning Model: Design and train a deep learning model that can effectively learn the voice characteristics from the training data. The model should be able to capture the nuances of the target speaker's voice and generalize well to unseeninput.

Real-Time Processing: Implement an efficient and optimized system that can perform voice cloning in real time, allowing for immediate conversion of the input speech to the synthesized voice output. The system should process the audio input in atimely manner to maintain a smooth and natural conversation flow.

Naturalness and Quality: Ensure that the synthesized voice output is highly natural and indistinguishable from the original speaker's voice. The system should preserve the speaker's individuality and produce high-quality, human-like speech.

Robustness and Generalization: Test and validate the system's performance on a diverse set of speakers, including different genders, ages, and accents. The system should be robust enough to handle variations in speech patterns and adapt to differentspeaking styles.

The ultimate goal of the project is to develop a reliable and accurate real-time voice cloning system that can be used in various applications, such as speech synthesis, virtual assistants, and voice dubbing, while ensuring the privacy and ethicaluse of the technology.

3. METHODOLOGY

3.1 Data Collection:

- Collect a dataset of audio recordings from the target speaker, containing a wide range of speech samples.
- Ensure the dataset includes various phonetic sounds, speaking styles, and emotions to capture the speaker's voice characteristics comprehensively.

3.2 Preprocessing:

- Clean and preprocess the audio recordings to remove noise, normalize volume levels, and standardize the format.
- Split the dataset into training, validation, and testing sets for model development and evaluation.

3.3 Speaker Encoder Training:

- Train a speaker encoder model using the training set.
- The model learns to extract speaker embeddings from input speech signals, capturing unique speaker characteristics.
- Use a loss function, such as triplet loss or contrastive loss, to encourage speaker embeddings from the same speaker to becloser while pushing apart embeddings from different speakers.

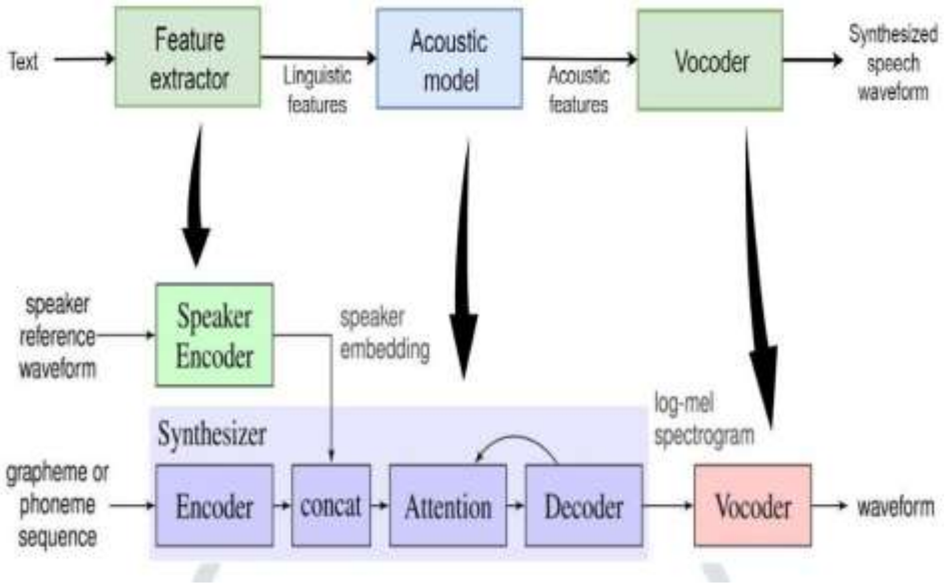
3.4 Synthesizer Training:

- Train a synthesizer model using the training set and the speaker embeddings obtained from the speaker encoder.
- The synthesizer generates mel-spectrograms from input text, conditioned on the speaker embeddings.

3.5 Vocoder Training:

- Train a vocoder model using the training set and paired mel-spectrograms and audio waveforms.
- The vocoder synthesizes high-quality audio waveforms from mel-spectrograms.

4. BLOCK DIAGRAM

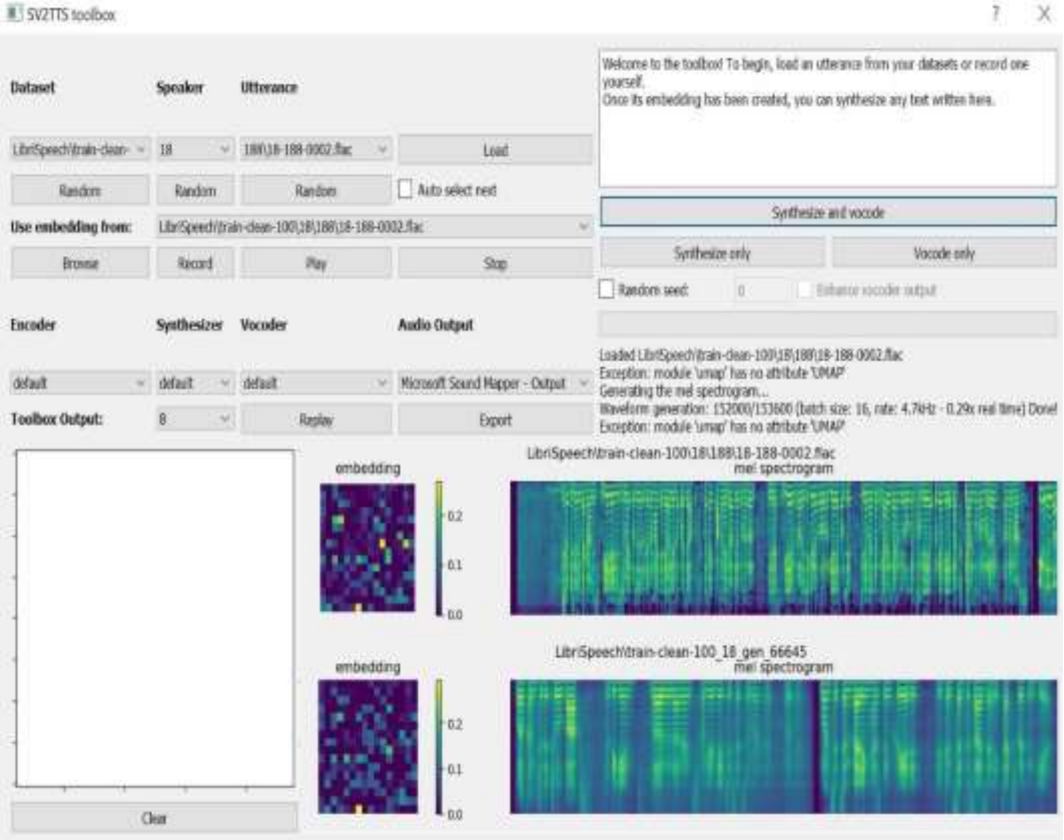


The above diagram elaborates the flow of the processes we plan to use for the execution of our project. The Encoder takes in the input audio and creates voice embeddings which have the characteristics of the unique speaker voice. The Synthesizer generates a grapheme or phoneme sequence from the input 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.

5. RESULT

Real-time voice cloning is the process of creating a synthetic voice that sounds like a particular individual in real-time. This is typically achieved through the use of machine learning algorithms such as deep neural networks that are trained on audio samples of the target voice.The result of a successful real-time voice cloning project would be a system that is able to capture the nuances of a person's voice and generate speech that is indistinguishable from the original speaker in real-time. Such a system could have a wide range of applications, including virtual assistants, chatbots, and even voice actors for video games and animated films.

This project has successfully developed a framework for real-time voice cloning that has not been published.



Dataset: LibriSpeech/train-clean-
Speaker: 18
Utterance: 188\18-188-0002.flac
Use embedding from: LibriSpeech/train-clean-100\18\188\18-188-0002.flac
Random seed: 0
Encoder: default
Synthesizer: default
Vocoder: default
Audio Output: Microsoft Sound Mapper - Output
Toolbox Output: 8

Welcome to the toolbox! To begin, load an utterance from your datasets or record one yourself.
Once its embedding has been created, you can synthesize any text written here.

```
Loaded LibriSpeech/train-clean-100\18\188\18-188-0002.flac
Exception: module 'umap' has no attribute 'UMAP'
Generating the mel spectrogram...
Waveform generation: 152000/153600 (batch size: 16, rate: 4.7kHz - 0.29x real time) Done!
Exception: module 'umap' has no attribute 'UMAP'
```

6. CONCLUSION

The Smart Waste Management System is a relatively powerful and reasonably priced solution for managing waste in plenty of settings, such as homes, offices, and public areas. The system's superior capabilities allow it to robotically locate the level and form of waste inside the bin and offer actual-time updates to customers through a mobile application. The use of ultrasonic sensors and servos to control the bin's lid and segregate waste by means of type reduces the need for human intervention and guarantees right waste disposal, contributing to a cleanser and healthier environment.

In addition, the combination of the Blynk IoT platform makes it possible to remotely monitor the machine, including to its comfort and person-friendliness. The machine's LCD show provides local updates at the bin's fame, which enables protection personnel to manage the waste useful to the society.

Overall, the Smart Waste Management System is a especially progressive solution which can help cope with the growing waste control challenges going through city regions global. By selling proper waste disposal and lowering the environmental effect of waste, this machine represents a considerable step in the direction of a greater sustainable future

The results aresatisfactory and the framework's ability to replicate speech is very good. 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. We believe that more powerful forms voice clones will be available in the near future.

7. REFERENCES

[1] Gilles Loupe, Corentin Jemine. Master Thesis: Automatic Multispeaker Voice Cloning. Faculty of Science Applications, University of Liege.

[2] Sander Dieleman, Heiga Zen, Aaron van den Oord, Karen Simonyan, Nal Kalchbrenner, Andrew W. Senior, Oriol Vinyals, Alex Graves and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. CoRR, abs/1609.03499, 2016

[3] Karen Simonyan, Seb Noury, Norman Casagrande, Nal Kalchbrenner, Erich Elsen, Edward Lockhart, Florian Stimberg,Sander Dieleman, and Koray Kavukcuoglu, Aaron van den Oord. Efficient neural audio synthesis, 2018.

[4] R. J. Skerry-Ryan, Rif A. Saurous, Jonathan Shen, Ruoming Pang, Ron J. Weiss,Yuxuan Wang, Yannis Agiomyrgiannakis, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, and YonghuiWu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. CoRR, abs/1712.05884, 2017.

[5] G. Penn and S. Shirali-Shahreza. Mos-naturalness and the quest for human-like speech. In 2018 IEEE Spoken Language Technology Workshop (SLT), pages 346 to 352, Dec 2018. doi:1109/SLT.2018.8639599.

[6] Andrew Gibiansky,Jonathan Raiman, John Miller, Sercan Arik, Gregory Diamos, Kainan Peng, Wei Ping, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech, 2017.

[7] Giuseppe Ruggiero, Enrico Zovato, Luigi Di Caro, Vincent Pollet. Voice Cloning: a MultiSpeaker Text-to-Speech Synthesis Approach based on Transfer Learning, 2021. REAL TIME VOICE CLONING Dept. of CSE, SaIT, Bangalore 46.

[8] Corentin James. Real Time Voice Cloning, 2019.

[9] Sercan O Arik, Jitong Chen, Kainan Peng, Wei Ping,Yanqi Zhou. Neural Voice Cloning with a few Samples, 2018.

[10] Jian Cong, Shan Yang, Lei Xie, Guoqiao Yu, Guanglu Wan. Data Efficient Voice Cloning fom Noisy Samples with DomainAdversial Training, 2020.