REALTIME VOICE CLONING USING DEEP LEARNING

[1]Dr. Madhusudhana G K, [2]Hruthik B Gowda, [3]Karun Datta Ramakumar, [4]Sheethal V, [5]Sushma M

[2,3,4,5]Student
Computer Science and Engineering
Vidya Vikas Institute of Engineering and Technology,
Mysore

***Abstract*- A Team of researchers has developed a novel method for voice cloning that can clone an unseen voice using just a few seconds of reference speech without retraining the model. This technique involves a three-stage pipeline and has been found to produce highly natural-sounding results. The researchers aim to replicate this model and make it open source for public use. They also plan to incorporate a new vocoder model to enable real-time voice cloning. In summary, the goal is to create a three-stage deep learning system that can clone voices in real-time. The voice cloning framework is based on a 2018 Google paper, and there is currently only one public implementation available besides the one being developed by the team. The system is capable of capturing a realistic representation of a voice from just a 5-second speech utterance in digital format, and can perform text-to-speech using any voice extracted from this process. The team aims to replicate each of the three stages of the model using their own implementations or open-source ones, and will train these models on large datasets of tens of thousands of hours of speech from several thousand speakers for several weeks or months. They will also develop appropriate pipelines for pre-processing the information. The ultimate goal is to create successful models of deep learning that can be used for voice cloning. The team aims to evaluate the strengths and drawbacks of the existing voice cloning framework based on the 2018 Google paper. Their main objective is to optimize the system for real-time performance, which means capturing and generating speech in less time than it takes to produce the speech itself. The framework is expected to have the ability to clone voices that it has never encountered during training, and generate speech from text it has not seen before. The team will focus on developing strategies to improve the speed and efficiency of the framework, while also ensuring that the quality of the cloned voices and generated speech remains high.**

Introduction:

Deep learning models have become increasingly popular in various fields of computational machine learning, including the synthesis of artificial speech from text, known as Text-to-speech (TTS). Since 2016, deep models have been developed that can generate more natural-sounding speech than conventional concatenative methods. Research efforts have been focused on improving the effectiveness and naturalness of these deep models, as well as training them in an end-to-end fashion. Inference on GPU has made significant progress, and models have been shown to produce near-human naturalness in the quality of the generated speech. However, measuring speech naturalness is still largely subjective, and some argue that the boundary of human nature has already been reached. Nonetheless, the correlation between subjective metrics and actual human speech indicates that there is still room for improvement in TTS technology. The goal of a single-speaker TTS model is not to clone a voice, but rather to create a fixed model that can incorporate new voices with minimal input. A common approach to clone new speakers is to condition a pre-trained TTS template, which has the ability to generalize voices. To clone a new speaker's voice, a common approach is to use a speaker encoder model to derive a low-dimensional embedding from a reference speech sample. This embedding is then used to condition a TTS template that has been trained to generalize the voice. By using this method, it is possible to create a fixed model that can integrate new voices with relatively little additional information, making it more data-efficient and computationally faster than training separate TTS models for each speaker. This method is more data-efficient, faster, and less computationally expensive than training a separate TTS model for each speaker. The length of reference speech determines the similarity of the voice produced to the speaker's true voice. The goal of a single-speaker TTS model is not to clone a voice, but rather to create a fixed model that can incorporate new voices with minimal input. A common approach to clone new speakers is to condition a pre-trained TTS template, which has the ability to generalize voices. In order to clone a new speaker's voice, a speaker encoder model is used to derive a low-dimensional embedding from the reference speech input. This embedding can then be used to condition a TTS template, which has been trained to generalize the voice, and produce the synthesized speech output in the desired speaker's voice. This method is more data-efficient, faster, and less computationally expensive than training a separate TTS model for each speaker. The length of reference speech determines the similarity of the voice produced to the speaker's true voice. This is done by using a low-dimensional embedding derived from a speaker encoder model that takes reference speech as input. This approach is generally more data-efficient and faster than training a separate TTS model for each speaker, as well as being computationally less expensive. However, the length of reference speech required to clone a voice can vary widely among different methods, ranging from just a few seconds to half an hour per speaker. This aspect plays a crucial role in determining the similarity of the generated voice to the speaker's true voice.

Methodology:

Deep Learning: Deep learning is a powerful subset of machine learning that enables computers to learn from vast amounts of data and make predictions or decisions based on that learning. It involves the use of artificial neural networks, which are modelled after

the structure of the human brain. Deep learning has transformed many areas of technology, including computer vision, natural language processing, and speech recognition. In particular, it has enabled breakthroughs in autonomous driving, allowing cars to detect and respond to their surroundings. It is also a key technology behind voice control in a variety of consumer devices, making it possible for users to interact with their devices in a more natural and intuitive way. With its ability to learn from large datasets and make increasingly accurate predictions, deep learning has the potential to revolutionize many aspects of our lives.

Design: To enable the voice cloning system to read any text with a desired voice, it requires two inputs - the text to be read and a sample of the voice that the text should be read in. This requires the system to have a comprehensive understanding of both the text and voice characteristics. To ensure ease of use for non-technical users, the system should have a user-friendly interface. The system can be divided into two main parts: training the neural network and integrating it into a flask application. The neural network should be trained to effectively clone new voices and generate natural-sounding speech, while the flask application will provide a user-friendly interface for users to input text and voice samples, and generate the desired output.

Voice Cloning

The goal of this project is to create a system that can generate speech in any speaker's voice from any given text input. This process involves two main steps: voice cloning and text-to-speech (TTS) synthesis. It is crucial to select the most appropriate approach to achieve high levels of naturalness and intelligibility in the final output. TTS systems are generally evaluated based on these two factors. There are two primary methods for accomplishing TTS conversion.

i) Concatenative approach:

○Utilizes high-quality audio samples
○ Limited by the availability and diversity of the data
○ Combines segments of different audio recordings to create new synthesized speech
○ Resulting speech is clean and clear, but lacks emotional expression and may not sound phonetically accurate
○ Generally intelligible, but may not sound entirely natural.

ii) **Parametric approach**:

The parametric approach is a method used in text-to-speech (TTS) synthesis that is less restrictive compared to the concatenative approach. It is a statistical approach that uses sound features like spectrum, frequency, and amplitude to generate speech. Unlike the concatenative approach, the parametric approach requires less data and is more robust. The method involves training a model to learn the relationship between linguistic and acoustic features of speech, allowing it to generate speech from text input. This approach is widely used in modern TTS systems and has proven to be an effective method for producing natural-sounding speech The Deep Learning approach is considered the best method for producing high-quality TTS synthesis output. Instead of generating audio waveforms in a time series format, the Deep Learning approach directly produces sound samples. This approach uses neural networks that are trained on large amounts of speech data to learn the patterns and characteristics of human speech. The advantage of this method is that it can produce more natural-sounding speech with less distortion and less reliance on pre-existing audio data. This project also uses the Deep Learning approach to synthesize speech, allowing for more

There are models that can be used here,

i) **Wave n**et: Wave net is a deep learning model used for generating raw audio waveforms. While it produces high-quality speech that is more natural and less robotic than other TTS methods, it requires an extremely high amount of data to train and is computationally expensive, which makes it difficult to implement in real-world applications.

ii) **DeepVoice**: DeepVoice is a neural TTS model that was developed by Baidu's Silicon Valley AI Lab (SVAIL) and is based on a sequence-to-sequence learning framework with attention mechanisms. The model has less data requirements compared to Wave Net but it still does not produce very satisfactory results in terms of naturalness and intelligibility of the synthesized speech

iii) **SV2TTS**: is a real-time voice cloning system that employs zero-shot learning, meaning it can learn to clone a new speaker's voice without any training examples. The system is composed of three deep learning models that are trained independently, allowing each model to be trained on different data sources, reducing the need for high-quality multispeaker data. This method is highly efficient and can generate high-quality synthesized speech with very low latency

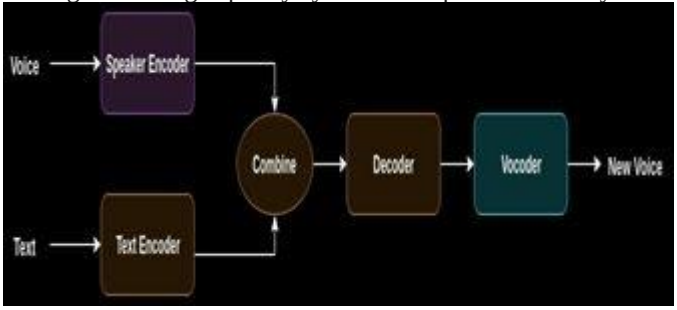

Fig:1.1 Working Flow of Diagram

REFERENCES:

[1] Jia, M., Zhang, S., Wei, L., Qin, T., & Liu, T. Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. arXiv preprint arXiv:1806

[2] Wang, Y., Ren, J., & Xu, B. (2019). Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6151-6155). IEEE.
[3] Zhang, C., Meng, Z., Wu, Y., & Wang, Y. (2020). Deep Voice Cloning: A Review. IEEE Access, 8, 146757-146772.
[4] Gibiansky, A., & Synnaeve, G. (2017). Deep voice 2: Multi-speaker neural text-to-speech. In 5th International Conference on Learning Representations, ICLR 2017-Workshop Track Proceedings.
[5] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Clark, R., & Battenberg, E. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In Advances in Neural Information Processing Systems (pp. 1187-1196).
[6] Chung, J. S., & Lee, S. Y. (2020). Fast Voice Cloning with Conditional Autoencoder and Residual Encoder-Decoder Networks. IEEE Access, 8, 151041-151056.
[7] Shen, J., & Wang, Y. (2019). Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. In ICASSP 2019-2019 IEEE. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7165-7169). IEEE.
[8] Yang, X., Wang, Y., Liu, J., & Xu, B. (2019).Generating high-quality speech from text with low latency using generative adversarial networks. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7100-7104). IEEE.
[9] Zhang, C., Wu, Y., Meng, Z., & Wang, Y. (2020). Voice cloning using speaker-adaptive training and generative adversarial networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 1509-1522.
[10] Song, X., Zhang, C., Meng, Z., Wu, Y., & Wang, Y. (2019). A fully convolutional neural network for real-time speech enhancement. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 246-250). IEEE.