REAL TIME VOICE CLONING USING DEEP LEARNING

Vidya Vikas Institute of Engineering and Technology

Dept. Of Computer Science and Engineering

Submitted by

Karun Datta Ramkumar (karundatta@gmail.com)

Hruthik. B. Gowda (hruthikdaali@gmail.com)

Sheethal (sheethal3031@gmail.com)

Sushma M (sushmam4321@gmail.com)

Under the Guidance of

Dr.Madhusudhan GK
Assistant professor
Dept of CSE, VVIET.

Introduction

A fascinating application of artificial intelligence (AI) is using Voice Cloning techniques that aim at precisely replicating someone's vocal style and characteristics. The term "cloning" here refers unequivocally not only about creating an exact replica but also with similar functionality and behavioural attributes. This cutting-edge technology utilizes machine learning models trained on vast audio datasets of a specific individual's voice to synthesize new speech that closely mirrors their unique vocal patterns.

Voice is the easiest and most natural mode of interaction for human beings. In computer science, cloning is the process of creating an exact copy of another application program or object. The term can be used to refer to an object, programming or an application that has similar functions and behaviour to another object or application program but does not contain the original source code from the concerned object or program.

Voice cloning has a variety of potential applications, including making it easier for people with speech impairments to communicate, creating more realistic-sounding virtual assistants, and even allowing actors to voice multiple characters in a single production. However, it also raises concerns about the potential for misuse, such as creating fake audio of public figures.

Objectives

- Develop a deep learning model that can accurately clone a person's voice in real-time using only a short audio sample as input.

- Explore and compare different deep learning architectures for voice cloning, such as convolutional neural networks (CNNs), and recurrent neural networks (RNNs), to determine the most effective approach.
- Improve the robustness of the voice cloning model by training it on a diverse range of voices, accents, and speech patterns to ensure that it can accurately clone a wide variety of speakers.

Methodology

The Recurrent neural network (RNN) help us to understand the cloning process and helps us to build our own cloning Model Using python This is a significant of finding because it means that voice cloning can be done without having to collect large amounts of data, which can be expensive and time-consuming. The AI is able to clone a voice by learning the patterns of vibration in the vocal cords that produce certain sounds. These patterns are then used to generate new similar sounds that can mimic the original voice.

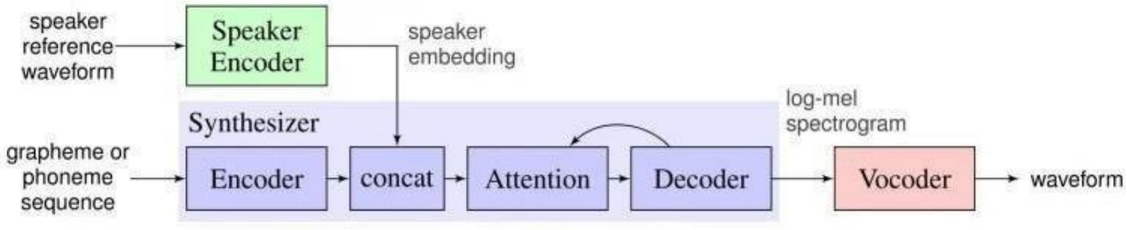

Fig: Architecture Diagram

The process shows recurrent speaker encoder, which computes a fixed dimensional vector, from a speech signal, followed by a sequence-to-sequence synthesizer, which predicts a Mel Spectrogram from a sequence of grapheme or phoneme inputs, conditioned on the speaker Embedding vector, and wavenet vocoder, which converts the spectrogram into time domain waveforms.

Speaker encoder :

- It is used to condition the synthesis network on a reference speech signal from the desired target speaker.
- Critical to good generalization is the use of a representation which captures the characteristics of different speakers, and the ability to identify these characteristics using only a short adaptation signal, independent of its phonetic content and background noise.

Synthesizer :

- It is trained on pairs of text transcript and target audio.
- The network is trained in a transfer learning configuration, using a pre trained speaker encoder to extract a speaker embedding from the target audio.

- No explicit speaker identifier labels are used during training.

Neural vocoder :

- We use the sample-by-sample autoregressive wavenet as a vocoder to invert synthesized Mel spectrograms emitted by the synthesis network into time domain waveforms.
- The architecture is composed of 30 dilated convolution layers. The network is not directly conditioned on the output of the speaker encoder. The Mel spectrogram predicted by the synthesizer network captures all of the relevant detail needed for high quality synthesis of a variety of voices.

Result and Conclusion

This project is Machine learning domain and the implementation is done using the high-level, general-purpose programming language Python. Our objective is to design a model för converting any text to any speaker's voice based on user selection in two steps i.e., Cloning the target voice and Text-to-Speech synthesis. We compared three models and found that SV2TTS matched our requirements. Though this we can understand that voice cloning is an area with a lot of possibilities of the technology being misused, we also cannot deny the fact that synthesizing text is a high technology advancement and artificial formation of speech given a text to be spoken.

.

Scope for future work

In future, we will further improve the quality of the speech by expanding the voices to all regional accents. We will improve the datasets size and quality. We will also work on the clarity of speech along with better consolidation of vocoder network.

Further we will work on the time consumption, storage space of the dataset and the accuracy of the system.