NEURAL VOICE CLONING

Team Members: Adil Ashique and Rosa Anil George

Abstract: Voice cloning is a highly desired capability for personalized speech interfaces. Neural network based speech synthesis has been shown to generate high quality speech for a large number of speakers. In this project, we aim to build a neural voice cloning system that takes only a few audio samples as input. Speaker adaptation technique is the one we would be using. Speaker adaptation is based on fine-tuning a multi-speaker generative model with a few cloning samples. In terms of naturalness of the speech and its similarity to original speaker, this method can achieve good performance with better naturalness and similarity, even with very few cloning audios.

Problem Statement - This project aims to build an application that will require few audio samples of the user as input and then using the speaker adaptation technique of voice cloning, will be able to read aloud a paragraph in the recorded voice.

Literature Survey

1. Paper on Neural Voice Cloning with few Samples - This is the research paper released by Baidu Research. This paper talks about two approaches for neural voice cloning: speaker adaptation and speaker encoding. Both approaches can achieve good cloning quality even with only a few cloning audios. For naturalness, both speaker adaptation and speaker encoding can achieve an MOS(Mean Opinion Score) for naturalness similar to baseline multi-speaker generative model.
2. Scaling Text to Speech with Convolutional Sequence Learning - This paper gives an insight to Deep Voice 3, a fully-convolutional attention-based neural text-to-speech (TTS) system. Deep Voice 3 matches state-of-the-art neural speech synthesis systems in naturalness while training an order of magnitude faster. We scale Deep Voice 3 to dataset sizes unprecedented for TTS, training on more than eight hundred hours of audio from over two thousand speakers.
3. A Robust Speaker-Adaptive HMM-based Text-to-Speech Synthesis -In this paper, the performance of several speech synthesis methods in unfavourable conditions are analysed. As a consequence of the investigations, a new robust training method for the speaker-adaptive HMM-based speech synthesis in for use with speech data collected in unfavourable conditions is proposed.
4. VCTK Dataset - This CSTR VCTK Corpus includes speech data uttered by 109 native speakers of English with various accents. Each speaker reads out about 400 sentences, most of which were selected from a newspaper plus the Rainbow Passage and an elicitation paragraph intended to identify the speaker's accent.

Objectives

To create an application that will :

1. Record the voice to be cloned
2. Convert any entered text to speech of the cloned voice.

Scope

This project is intended to be a long term project. Once we are able to achieve all the goals that are set in the project and it is able to use the cloned voice to read a document we supply it with we intend to further extend the application in two ways:

1) To serve as a chatbot that will converse with you in your voice or in a voice that you intend it to speak. This kind of application can be of great help in the field of mental therapy and in the use of personal health assistant applications.

2)The paper that we have cited have suggested that if the model is trained with samples of musical notes instead of human voice samples it will be able to generate music of its own. We would like to explore this option also further down the line.

Timeframe

Approx time of review	Expected Status
Phase One (Sept end)	Learning Phase to understand the concept of sequential neural network, the basics of voice cloning and to look into the various methods to convert text to speech.
Phase Two (Nov mid)	Learning Phase to have a proper understanding of the various methods of voice cloning that can be implemented and to choose the method that seems to be the most feasible one and gives the best results.
Phase Three (Jan end)	We plan to complete the following two objectives in the following time frame: 1)Train a model that can learn a voice and replicate it from a minute or so of voice sample 2)Use the voice that is generated by the model to read the text that has been supplied by the user.
Phase Four (Mar end)	Design the app that is required and to finalise how the transfer of data between the model and app will take place.

Project Budget

Access to GPUs for training the model.

Monitoring and Evaluation

Phase 1 will be marked by: i) KSS and hackathons that are conducted in order to deepen the understanding of the basics of neural nets and sequential neural nets.
ii) A few sessions will be held to understand the underlying concepts in the papers that we are using as reference for the project.

Phase 2 will be marked by: A method must zeroed in on after the study of the feasibility and limitations of the various methods hat we have considered. This can done after trying out the various approaches that we have seen through our literature survey and thus choose the one that delivers the most conclusive results.

Phase 3 will be marked by: We will have to train the model by this phase and it must be capable of replicating the voice that we want to clone from the voice sample that we have recorded for a minute or so.

Phase 4 will be marked by: A method to use the voice that we have cloned to convert the text that we have supplied to speech must be created by this phase. There are a few few github repositories on the text to speech converts and taking them as reference a method will be deployed to achieve the same.

By the final evaluation the model will be integrated into an app that will act as the interface for the model to take input from the user (which here will be the voice to be cloned and the text to be read) and give the desired output which is the text read out in the cloned voice.