Team: Tacocat

Project Title: Voice Cloning using Speech Embedding

Project Summary:

Development in text-to-speech technology opened doors to many interesting applications, such as voice assistants and text readers that we interact with every day. However, they all support very limited voices out of the box, as it is generally very challenging to acquire many hours of training data from a speaker to make it possible.

However, recent progress in the field allowed for custom voice cloning from a few audio samples, which can have multiple useful applications, such as helping those with vocal disabilities feel more comfortable with their alternative voice, generate audio for animations, or videos to speed up development, audio augmentation to supplement other speech models, and many more. We wanted to implement this process in an end-to-end fashion, while learning more about working with speech data.

Approach:

Baseline neural text-to-speech (TTS) is typically performed with two parts: a text to spectrogram decoder, and a spectrogram to waveform vocoder. Existing research on voice cloning extends this approach by adding a speaker encoder which produces a speaker embedding from a few voice samples to be used as an additional input to the spectrogram decoder.

For our project, we aim to add a speaker encoder to a traditional two-part TTS system. As a baseline we expect to replicate the speaker encoder from [1] and plan to use Tacotron2 and WaveGlow for the TTS stack. In order to keep training-time reasonable, we will look to use pre-trained weights for Tacotron and fine-tune with the speaker embedding instead of training from scratch.

As a stretch goal, we want to experiment with different speaker encoders in order to improve performance. One approach would be to use different architectures or objective functions to train the speaker encoder. We will attempt to evaluate the experiments using mean observer score (MOS).

Resources/Related Work:

The typical approach in related work is described above. More recent research has included: applying attention over speaker inputs rather than fixed-length speaker embeddings [4], varying methods of generating speaker embeddings [5] and network architectures which combine the TTS parts into an end-to-end system [6].

[1] "Neural Voice Cloning with a Few Samples", Arik et al.

[2] "Voice Imitating Text-to-Speech Neural Networks", Lee et al.

[3] "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis ", Jia et al.

[4] "Attentron: Few-Shot Text-to-Speech Utilizing Attention-Based Variable-Length Embedding", Choi et al.

[5] "Voice Cloning: a Multi-Speaker Text-to-Speech Synthesis Approach based on Transfer Learning", Ruggiero et al.

[6] "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech", Kim et al.

Datasets:

LibriTTS – http://www.openslr.org/60/ (annotated voice samples from audiobooks)

VCTK - https://datashare.ed.ac.uk/handle/10283/2950 (newspaper voice clips)

Team Members:

Seongmin Youm

Perry Chu

Nolan Piland

James Liu