VOICE CLONING: A MULTI-SPEAKER TEXT-TO-SPEECH SYNTHESIS APPROACH BASED ON TRANSFER LEARNING

Avinash Ramesh

Text-to-Speech (TTS): *A system that converts normal language text on a computer into audible speech output.*

Multi-Speaker TTS: Synthesizing speech with different voices with a single model.

Transfer learning (TL): It is a research problem in machine learning (ML) that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem

Model Architecture

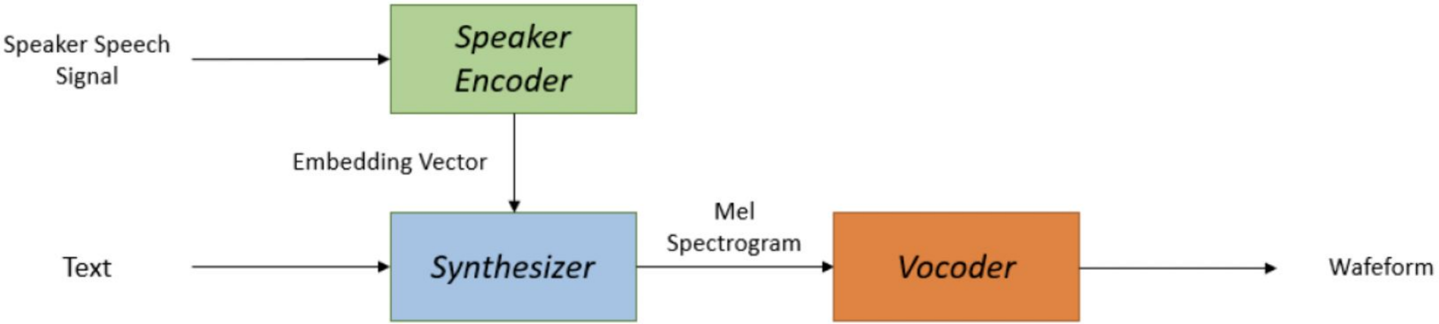

Fig. 1: High level overview of the three components of the system.

Speaker Encoder generates a fixed-dimensional embedding vector from a few seconds of reference speech from a target speaker

Synthesizer predicts a mel spectrogram[2] from an input text and an embedding vector

V**ocoder** infers time-domain waveforms from the synthesizer's mel spectrograms.

Baseline System - Corentin Jemine’s real-time voice cloning system

- a recurrent speaker encoder with three LSTM layers and a final linear layer
- each with 256 units
- a sequence-to-sequence with an attention synthesizer and WaveRNN as a vocoder.

Speaker Encoder: Proposed System

rec_conv network: 5 Conv1D layers, 1 GRU layer and a final linear layer

rec_conv_2 network: 3 Conv1D layers, 2 GRU layers each followed by a linear projection layer

gru network: 3 GRU layers each followed by a linear projection layer

advanced_gru network: 1 Conv1D layer and 3 GRU layers each followed by a linear projection

layer ***lstm network***: 1 Conv1D layer and 3 LSTM layers each followed by a linear projection layer

Table 1: Speaker Verification Equal Error Rates.

Name	Step Time	Train Loss	SV-EER	LR Decay
rec_conv	**0.33s**	0.36	0.073	Reduce on Plateau
rec_conv_2	**0.45s**	0.49	0.075	Reduce on Plateau
gru	1,45s	0.33	0.054	Every 100,000 step
advanced_gru	0.86s	**0.14**	**0.040**	Exponential
lstm	1.08s	0.17	0.052	Exponential

Similarity Evaluation

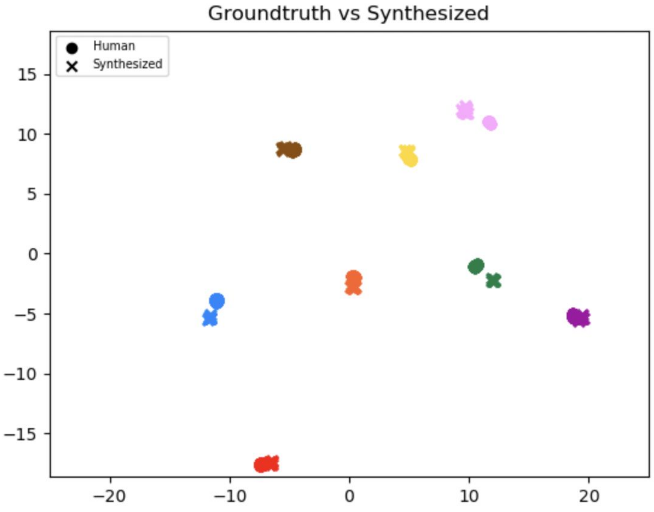

Fig. 5: Groundtruth utterance embeddings vs the corresponding generated ones of the 8 speakers chosen for testing.

Subjective Evaluation

Table 2: MSS of the baseline and the proposed systems.

System	MSS
baseline	2.59 ± 1.03
proposed	3.17 ± 0.97

Conclusion

- The author’s goal was to create a voice cloning system that could generate natural speech for a variety of target speakers while using minimal data.
- Their system combines a speaker encoder network that has been trained independently, a sequence-to-sequence with attention architecture, and a neural vocoder model.
- The synthesizer and vocoder can generate good-quality speech even for speakers who have never been observed before by using a transfer learning technique.
- Despite the experiments demonstrating a reasonable similarity to real speech and improvements over the baseline, the proposed system falls short of human-level naturalness when compared to single-speaker results

Thank You!