CS230

Multispeaker Speech Synthesis with Configurable Emotions

Jiaxiang Fu
jiaxiang@stanford.edu

1 Introduction

Most of widely used speech synthesis systems are only able to generate voices of a limited number of speakers, and they usually require large amount of voice samples from these speakers. Furthermore, they can only generate monotonous speeches with very little variations in tone or emotion.

In this project we propose a way to condition our speech synthesis model on both speaker voice and configurable emotions. Given an input text, the model should generate the corresponding speech audio with a voice similar to the speaker in a given reference audio, and convey an emotion either of a given category or similar to another reference audio. Ideally, the model should work reasonably well with reference audios shorter than 30 seconds.

2 Related work

Tacotron [3] introduced the first end-to-end model for speech synthesis trained directly on text-audio pairs. This allowed us to avoid any hand-crafted feature representations. Tacotron 2 [4] improved upon Tacotron by using WaveNet [5] as its vocoder to generate human-like natural speeches. Importantly, the encoder-decoder seq2seq architecture used in Tacotron 2 made it easy for us to alter the encoding to generate speeches with different styles.

Jia et al. [1] applied transfer learning from speaker verification to generate speeches that mimic the voices of different speakers. They do so by concatenating the speaker embeddings produced by pre-trained speaker verification model with the encodings produced by Tacotron 2, and fed the concatenated representation back to the decoder and vocoder of Tacotron 2. However, the model does not support varied emotions.

It has been shown that embeddings can also be used to condition the Tacotron decoder to generate speech with different prosody styles [6, 7]. Based on that, Um et al. [2] trained embeddings that encode the emotions of speeches. Further, they proposed a linear interpolation method to control the intensity of emotions in the synthesized speech. This model only supports a single speaker.

In this work we plan to apply apply both the speaker embedding method in [1] and emotion control in [2].

3 Dataset

There are plenty of readily available multispeaker speech synthesis datasets. Notably, VCTK [8] has 44 hours of speech from 109 speakers. LibriSpeech [9] has 436 hours of speech from 1,172 speakers.

For emotion encoding, the primary dataset for this work is SAVEE (Surrey Audio-Visual Expressed Emotion). However, with only 480 utterances covering 7 emotions, this dataset is potentially not be

CS230: Deep Learning, Fall 2021, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

sufficient for training a good model. If necessary, additional dataset may be used or sourced from online platforms such as Youtube.

4 Methods

We will use an end-to-end architecture with 5 main components similar to [1, 2, 6, 7]. The 5 components are (1) text encoder, (2) speech synthesizer, (3) vocoder, (4) speaker encoder and (5) emotion encoder. The output of all 3 types of encodes are then concatenated and used by the synthesizer and vocoder.

For training we apply transfer learning whenever possible by fune-tuning pre-trained components with optionally frozen layers. Particularly, we plan to use triplet loss to train the emotion encoder, before all components are combined together for end-to-end training.

5 Evaluation Metrics

The primary metric we use is the Mean Opinion Score (MOS) which is also used in many related work [1, 2, 3, 4, 5, 6]. We will source human listeners primarily from peers in the CS230 class.

To ensure fair comparison, we present to the human listeners pairs of audios produced by our model and the baseline model without labels and in random orders. Where applicable we also include the ground truth to form a trio that human listeners can compare side by side.

In total we evaluate 4 different aspects of the resulting audios - correctness, naturalness, similarity to target speaker and emotion richness.

References

[1] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno and Yonghui Wu. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. Advances in Neural Information Processing Systems 31 (2018), 4485-4495

[2] Se-Yun Um, Sangshin Oh, Kyungguen Byun, Inseon Jang, ChungHyun Ahn and Hong-Goo Kang. Emotional speech synthesis with rich and granularized control. ICASSP 2020-2020 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)., pp. 7254-7258, 2020.

[3] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis. In Proc. Interspeech, pages 4006–4010, August 2017.

[4] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui. Wu. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018.

[5] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. CoRR abs/1609.03499, 2016.

[6] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, et al., "Style tokens: Unsupervised style modeling control and transfer in end-to-end speech synthesis", Proc. ICML, 2018.

[7] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, et al., "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron", Proc. ICML, 2018.

[8] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit, 2017.

[9] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: an ASR corpus based on public domain audio books. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, pages 5206–5210. IEEE, 2015.