OSM: One-Shot Multi-Speaker Text-to-Speech

Nikolay Kozyrskiy, Gleb Balitskiy

Status Report №1, April 2021

1 Problem Statement

One-Shot Multi-Speaker Text-to-Speech (OS MS TTS) systems are aimed to transform text into speech with voice determined by small single sample. The main problem here is to reproduce the new unseen voice without retraining the network. There is an approach with three main stages which is used to solve this problem. The unique for each voice speaker embeddings, which reveal the voice characteristics, are generated at the first stage (*Speaker Encoder*). At the second stage (*Synthesizer*) the text is transformed to mel-spectrogram using previously obtained embeddings. Finally, the speech is reproduced from the mel-spectrogram with the *Vocoder*. But there is lack of implementations with these three parts properly combined. So the goal of our project is to create a flexible framework to combine these parts and provide replaceable modules and methods in each part.

2 Main Challenges

By now we see the following main challenges:

- The solution to our problem consists of three subtasks, which already have a great solutions. Therefore, the existing solutions for OS MS TTS are essentially a compilation of solutions for these individual problems, for which there are many ready-made and well-implemented solutions. The main challenge is to made the framework flexible and ensure the compatibility of individual parts.
- The methods used in each subtask differ in the set of parameters and the nature of the algorithm. Therefore, it will be quite difficult to provide a single API.

3 Baseline Solution

We choose solution proposed by the instructors as a baseline, which can be found here. It is the implementation of [1] made in Google in 2018. Here authors

use the speaker encoder, presented in [2], which generates a fixed-dimensional embedding vector known as d-vector. As for Synthesizer they use model based on Tacotron 2 [3] while an auto-regressive WaveNet-based is used as the Vocoder [4].

3.1 Pros and Cons

The Real-Time-Voice-Cloning contains the realizations of encoder, Tacotron 2 and WaveRNN. The whole pipeline described in [1], including preprocessing steps, is also implemented in this repository. However, the project is not flexible enough. More specifically, in the current state it cannot be used as the framework for One-Shot Multi-Speaker Text-to-Speech system as there are no convenient mechanisms for manipulating with the three main modules. For example, the proposed multi-speaker TTS system in [5] cannot be easily implemented with the help of Real-Time-Voice-Cloning as there are no extensibility points which allow to adjust the pipeline for the new method.

3.2 Our Improvement

Our plan is to use the Real-Time-Voice-Cloning as starting point with implemented baseline. We will introduce the flexible modular design of the framework. Such approach will help us to create the convenient API for external users who will be able to use our framework for incorporating the Multi-Speaker TTS system in their products. The API will also let the users customize modules and pipeline steps without changing the source code of the framework if needed. We will implement several Speaker Encoders (LDE, TDNN) and add them to our framework as well.

4 Roles of the Participants

Nikolay will design the modular architecture, API for external usage and training pipeline. Gleb will implement working stack of models, write documentations and usage examples.

5 Link to the GitHub repository

All the materials are available at OSM-one-shot-multispeaker.

References

[1] Ye Jia, Y. Zhang, Ron J. Weiss, Q. Wang, Jonathan Shen, Fei Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez-Moreno, and Y. Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *NeurIPS*, 2018.

[2] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883, 2018.

[3] Jonathan Shen, R. Pang, Ron J. Weiss, M. Schuster, Navdeep Jaitly, Z. Yang, Z. Chen, Yu Zhang, Yuxuan Wang, R. Skerry-Ryan, R. Saurous, Yannis Agiomyrgiannakis, and Y. Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, 2018.

[4] Aäron van den Oord, S. Dieleman, H. Zen, K. Simonyan, Oriol Vinyals, A. Graves, Nal Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. In *SSW*, 2016.

[5] Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6184–6188, 2020.