IE643: Deep Learning : Theory and Practice **Aug-Dec 2020**

End-term Project Report : Transfer Learning from Speaker Verification to Multispeaker TTS Synthesis

Team Name: Google2 *Team Members: 170050005, 170050031, 170050050*

Abstract

This project report contains modifications to the original paper which proposed a neural network-based text-to-speech synthesis for multiple speakers, even those unseen during the training. The system consists of three independently trained components - (1) a speaker encoder network, trained on a speaker verification task to generate a fixed-dimensional embedding vector from only seconds of reference speech from a target speaker; (2) a sequence-to-sequence synthesis network based on Tacotron 2 that generates a mel-spectrogram from text, conditioned on the speaker embedding; (3) an auto-regressive WaveNet-based vocoder network that converts the mel-spectrogram into time domain waveform samples. The proposed model is able to transfer the knowledge of speaker variability learned by the discriminatively-trained speaker encoder to the multispeaker TTS task, and is able to synthesize natural speech from speakers unseen during training. Finally, we propose modifications to the existing architecture to obtain speech that is more natural and yet distinguishable at the same time, than the one generated by the original model.

1 Introduction

The goal of this project was to implement and possibly improve upon the existing architecture proposed by Zhang et. al. [1] in their paper. The said paper aims to built a TTS system which can generate natural speech for a variety of speakers in a data efficient manner. It specifically addresses zero-shot learning setting, where a few seconds of untranscribed audio from a target speaker is used to synthesize speech in that speaker's voice without updating any model parameters.

Systems with zero-shot learning have accessibility applications, such as restoring the ability to communicate naturally to users who have lost their voice and are therefore unable to provide many new training examples. They could also enable new applications, such as transferring a voice across languages for more natural speech-to-speech translation, or generating realistic speech from text in low resource settings. To prevent the potential misuse of the this technology, the paper proposed a verification method to distinguish between the real and the generated voice.

Synthesizing natural speech requires training on a large number of high quality speech-transcript pairs. This is what makes it impractical. The paper proposed to decouple speaker modeling from speech synthesis by independently training a speaker-discriminative embedding network that captures the space of speaker characteristics and training a high quality TTS model on a smaller dataset conditioned on the representation learned by the first network. Decoupling the networks enables them to be trained on independent data.

We provide a survey of existing literature in Section 2. Our proposal for the project is described in Section 3. We give details on experiments in Section 5. A description of future work is given in Section 7. We conclude with a short summary and pointers to forthcoming work in Section 8.

2 Literature Survey

There has been significant interest in end-to-end training of TTS models, which are trained directly from text-audio pairs, without depending on hand crafted intermediate representations like [9].

DeepMind introduced a new parametric TTS called WaveNet [5] which uses stack of casual convolutions that have various dilation factors. It produces great quality human like audio but required complex inputs like linguistic features, fundamental frequencies and phoneme durations.

Tacotron [6] model was introduced that only took raw character data as input to generate speech. It produced magnitude spectrogram which was further converted to speech using Griffin-Lim vocoder. Tacotron 2 used WaveNet as a vocoder to invert spectrograms generated by an encoder-decoder architecture with attention. It aimed at obtaining naturalness approaching that of human speech by combining Tacotron's prosody with WaveNet's audio quality. The only downside was that it only supported a single speaker.

Gibiansky et al. [3] introduced a multispeaker variation of Tacotron in the DeepVoice2 architecture. The model learned low-dimensional speaker embedding for each training speaker. Deep Voice 3 [4] proposed a fully convolutional encoder-decoder architecture which scaled up to support over 2,400 speakers from LibriSpeech. These systems learn a fixed set of speaker embeddings and therefore only support synthesis of voices seen during training.

In contrast, VoiceLoop [10] proposed a novel architecture based on a fixed size memory buffer which can generate speech from voices unseen during training. Obtaining good results required tens of minutes of enrollment speech and transcripts for a new speaker. Recent extensions have enabled few-shot speaker adaptation where only a few seconds of speech per speaker (without transcripts) can be used to generate new speech in that speaker's voice. Another work extends Deep Voice 3, comparing a speaker adaptation method similar to [10] where the model parameters (including speaker embedding) are fine-tuned on a small amount of adaptation data to a speaker encoding method which uses a neural network to predict speaker embedding directly from a spectrogram. The latter approach is significantly more data efficient, obtaining higher naturalness using small amounts of adaptation data, in as few as one or two utterances. It is also significantly more computationally efficient since it does not require hundreds of backpropagation iterations.

Nachmani et al. [11] similarly extended VoiceLoop to utilize a target speaker encoding network to predict a speaker embedding. This network is trained jointly with the synthesis network using a contrastive triplet loss to ensure that embeddings predicted from utterances by the same speaker are closer than embeddings computed from different speakers. In addition, a cycle-consistency loss is used to ensure that the synthesized speech encodes to a similar embedding as the adaptation utterance

3 Methods and Approaches

The proposed architecture consists of three components - (1) a recurrent speaker encoder to compute a fixed dimensional vector containing just the speaker characteristics from a speech signal (2) a sequence-to-sequence synthesizer which predicts a mel spectrogram from a sequence of grapheme or phoneme inputs (3) a autoregressive WaveNet vocoder to convert the spectrogram into time domain waveforms

Speaker encoder uses a speaker-discriminative model trained on a text-independent speaker verification task. It is based on Generalised End-to-End Speaker Verification [7]. The input is a 40-channel log-mel spectrogram, passed on to a stack of 3 LSTM layers of 768 cells, each followed by a projection to 256 dimensions. The final embeddings is created by L2-normalizing the output of the top layer's final frame.

Synthesizer extends Tacotron2 architecture to support multiple speakers as in DeepVoice2. The input is a

character sequence with predicted mel-spectogram frames as output. The model augments Tacotron2's L2 loss with an additional L1 loss. Transfer learning is used here without any changes to the encoder parameters.

Vocoder consists of autoregressive WaveNet vocoder as defined in Tactron2 architecture with 30-dilated convolutional layers. Is uses mel-spectogram frames from the synthesizer as input with time-domain waveforms as output.

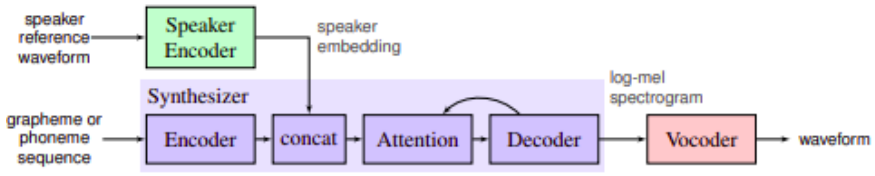

Figure 1: Model overview

3.1 Work done before mid-term project review

Prior to mid-term project review, the team looked into the details and background of the of the assigned paper. We reviewed at the related papers like DeepVoice and Tacotron and tried to understand the architectures involved. We looked for the existing codebase available that is related to the paper. Fortunately, there was existing code for the paper. We cloned the code and tried to understand its structure and working. We experimented with the code. Following this, we tried to follow up with the experiments mentioned in the paper to get the desired results. There was a flaw in the paper training process. We suggested a possible way of improvement and empirically justified that the new technique performs better.

3.2 Work done after mid-term project review

After the mid-term review, we started with setting up the GPU to compare the suggested improvement and evaluate the results against the existing observations of the paper. However, considering the fact that it would require a lot of days of training, we skipped the part and used the pre-trained models available for further training. We proceeded with formulating the modifications to the existing work and trying to compare the results with the pre-trained models. We used Surrey Audio-Visual Expressed Emotion(SAVEE) [16] to incoporate emotions into speech. Further, details are provided in section 5.

4 Dataset Details

The three components of the architecture are trained independently. VCTK contains 44 hours of clean speech from 109 speakers, the majority of which have British accents. LibriSpeech [12] consists of the union of the two "clean" training sets, comprising 436 hours of speech from 1,172 speakers, sampled at 16 kHz. Other publicaly available datasets used are LibriTTS [13], VoxCeleb [14] and VoxCeleb2 [14]. Anglophone nationalities are used for the VoxCeleb datasets. For further trainig the model for emotions, we used Surrey Audio-Visual Expressed Emotion(SAVEE) [16] dataset.

For the data preprocessing, the audio is downsampled to 16 kHz, trimmed leading and trailing silence (reducing the median duration from 3.3 seconds to 1.8 seconds), and split into three subsets: train, validation and test sets as available on the websites.

5 Experiments

For evaluating our modification proposed in midsem, we had to train a new synthesizer. This training was CPU intensive, we were provided with SSH access to an external Nvidia 1080 GPU. It has 8GB DDR5 RAM with 256-bit bus width and 1.6GHz base clock frequency. To further reduce the computation time needed, we re-trained the available pre-trained model instead of training it from scratch and we trained this model only on one percent of the entire dataset. We, then, trained the existing models for about 68000 iterations on the above mentioned datasets for 11 days. We tried to use a combination of utterances from the same speaker for better capturing of the speaker embeddings. This yielded results but not with much success. The results were, however, inconclusive.

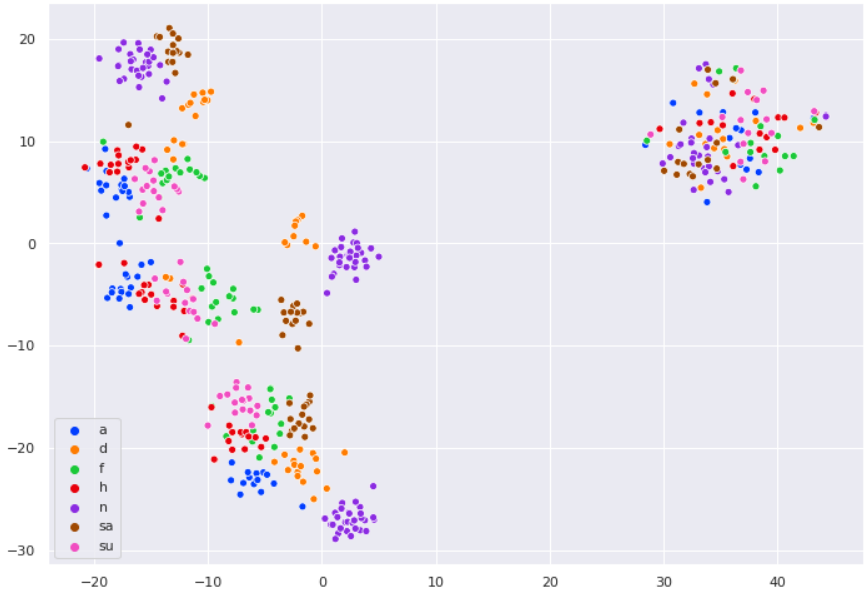

Figure 2: Emotion separation using tSNE (t-distributed Stochastic Neighbor)

As stated on the LibriSpeech website, All of the data comes from audiobooks. Thus, the audios are monotonous and lacks emotions. This is where propose the modifications to make the speech more realistic. We found Surrey Audio-Visual Expressed Emotion (SAVEE) [16] dataset, consisting of four speakers with their audio samples in different emotions. Using model encoder, we showed that two person speaking in same emotion are closer in embedded represention than two different emotions by same person. This implies that same emotions vector cluster together, thus we can calculate a vector direction that points to a particular emotion cluster, we then add this is emotion vector to speaker embedding before generating speech.

6 Results

The evaluation of the project primarily relies on crowdsourced Mean Opinion Score (MOS) [8] evaluations based on subjective listening tests. All the MOS evaluations are aligned to the Absolute Category Rating scale, with rating scores from 1 to 5 in 0.5 point increments. This framework is used to evaluate synthesized speech along two dimensions: its naturalness and similarity to real speech from the target speaker. Because this method is based on crowd sourcing, it is not possible to replicate the results.

For our first experiment, we found that almost similar audio files are generated. The reason might be only

a fractional epochs were run on a small portion of data. Overall, result of the first experiment is inconclusive. In second experiment, instead of MOS, to test out the idea of combining emotions into the speaker embeddings turned out pretty good. We selected three samples randomly. The pre-trained model and the one with modifications was used to obtain out-put speech samples. A survey form was floated on MS Teams group. So far the responses received indicate that the use of emotions have boosted the speech quality in two out of the three samples. The idea of combining emotions into the speaker embeddings turned out pretty good.

7 Future Work

Future modifications to the paper can be made in the directions of adding emotions to the speech to make it look more natural. In the project, we have introduced the concept of happiness vector being added to the speaker embeddings. Continuing this work, other emotions like anger and sadness can be merged to match the naturalness of human speech. Proper training with larger datasets and modifications can show much better results than those obtained in this project. Future work should aim towards achieving distinguishable speech based on the emotions.

8 Conclusion

The project aimed at implementing an existing research paper for TTS for multiple speakers, even those unseen during training with zero-shot learning. The paper itself has done a great work in this field by combining different state of the art architectures like Tacotron2 and WaveNet. The results obtained are quite substantial in the field of speech synthesis in a plethora of applications. We proposed modifications to the existing framework by adding emotions to the speech. This makes the synthesized speech more realistic. Directions have been provided above for future work in this field.

References

[1] *Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis, 2019* Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu

[2] *Deep voice: Real-time neural text-to-speech, 2017* S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi

[3] *Deep Voice 2: Multi-Speaker Neural Text-to-Speech, 2017* Sercan Arik and Gregory Diamos and Andrew Gibiansky and John Miller and Kainan Peng and Wei Ping and Jonathan Raiman and Yanqi Zhou

[4] *Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning, 2018* Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, John Miller

[5] *Wavenet: A generative model for raw audio, 2016* A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu

[6] *Tacotron: Towards end-to-end speech synthesis, 2017* Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio

[7] *Generalized end-to-end loss for speaker verification, 2018* Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno

[8] *Crowdmos: An approach for crowdsourcing mean opinion score studies, 2011* F. Ribeiro, D. Florêncio, C. Zhang, and M. Seltzer

[9] *Char2Wav: End-to-end speech synthesis, 2017* Jose Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio

[10] *VoiceLoop: Voice fitting and synthesis via a phonological loop, 2018* Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani

[11] *Fitting new speakers based on a short untranscribed sample* Eliya Nachmani, Adam Polyak, Yaniv Taigman, and Lior Wolf

[12] LibriSpeech dataset available at `http://www.openslr.org/12`

[13] LibriTTS dataset available at `http://www.openslr.org/60`

[14] VoxCeleb dataset available at `https://www.robots.ox.ac.uk/ vgg/data/voxceleb/`

[15] VoxCeleb2 dataset available at `https://www.robots.ox.ac.uk/ vgg/data/voxceleb/vox2.html`

[16] SAVEE dataset available at `http://kahlan.eps.surrey.ac.uk/savee/`