Conditioned Text-To-Speech System

Martin Iglesias Goyanes
KTH Royal Institute of Technology
martinig@kth.se

Jakub Reha
KTH Royal Institute of Technology
jakubr@kth.se

Abstract

Our original plan was to reproduce the work in "Controllable Emotion Transfer For End-to-End Speech Synthesis" [1], which was done on a single Chinese speaker. In our setting we used the multi-speaker English IEMOCAP dataset, which means the system had to be adapted. We modified an already existing implementation [2] of Tacotron2 [3] from Nvidia and used already pre-trained vocoder WaveGlow [4]. After initial attempts we realized that the IEMOCAP dataset lacks in the amount and quality of data for our purpose. Therefore, we pivoted our project and focused on 3 different approaches to condition Tacotron2 on only the speaker's identity and train on the VCTK dataset. This was done successfully and it was even showed that the approach can be used on unseen speakers.

1 Introduction

Since its inception one of the main goals of text-to-speech (TTS) has been to provide clear intelligible speech at or around human level. However, speech is not only about what is being said, but also how it is being said. And being able to control the synthesized speech has many business applications- Siri, Alexa, aids for blind or mute people, E-learning, audiobooks, automatic translation, telecommunication, GPS .... With the advent neural network based speech synthesis systems the goal of producing intelligible speech have more or less been achieved, and thus the goal has been moved to trying to control every specific aspect of how the speech sounds. Some of these aspects are the speakers identity and speakers emotion. In this project we explore Tacotron2 TTS synthesis conditioned on these two aspects.

TTS is a huge field with many approaches. [5] presents a very nice overview of these. We are aware that there are systems that claim to be faster [6; 7] or even to produce better quality speech than Tacotron2. However, we still see Tacotron2 as the standard TTS system as new methods are building upon its architecture [1; 8] and others [6; 7; 9] still compare their results to Tacotron's. Therefore, we feel it is important to understand how Tacotron works.

2 Method

2.1 IEMOCAP Dataset

2.1.1 Description

The data used in the emotion transfer experiments of this project was the Interactive Emotional Dyadic Motion Capture (IEMOCAP) data set. It contains around 12 hours of English spoken dialog from 10 speakers (5 male and 5 female), half of the data is scripted and the other half is improvised. It is composed of video recordings, motion captures of the speakers face, transcriptions, expressed emotions and more. The recordings are separated into 5 conversations (Sessions) with one female speaker and one male speaker each, therefore speakers ids have the form: Ses1F, Ses1M, ... The utterances in the data set have been annotated by at least 3 people. The data we are using in this study are only the audio recording, transcript, speaker identity and emotion. The emotion of the speaker are categorized into one of the following categories: anger, happiness, excitement, sadness, frustration, fear, surprise, other and neutral state. However, not all utterances are categorized or there is no majority agreed category. The classes are also heavily imbalanced.

2.1.2 Preprocessing

The audio from the IEMOCAP [cite] data set was also preprocessed before being fed to the modified Tacotron2 model. The audio files were resampled to 22kHZ, any silence in the beginning or end of the sample were removed as we found this to be detrimental for the Tacotron2 training. Utterances belonging to categories "fear", "other" and utterances with non-agreed emotional labels were removed. This left us with 6.4 hours of clean preprocessed data. Since the classes were still imbalanced we also computed inversely proportional weights for each label that would balance the loss computations.

2.2 VCTK Dataset

2.2.1 Description

The employed CSTR VCTK Corpus [10] contains voice data from 110 English speakers with different accents. Each speaker reads around 400 phrases from a newspaper, as well as the rainbow passage and an elicitation piece for the speech accent archive. With permission from Herald & Times Group, the newspaper texts were stolen from Herald Glasgow. A greedy algorithm selects a distinct collection of newspaper texts for each speaker, increasing the contextual and phonetic coverage. The details of the text selection algorithms are described in https://doi.org/10.1109/ICSDA.2013.6709856. All speakers read out the same rainbow passage and elicitation paragraph.

All speech data was captured in a hemi-anechoic room at the University of Edinburgh using the same recording setup: an omni-directional microphone (DPA 4035) and a tiny diaphragm condenser microphone with very broad bandwidth (Sennheiser MKH 800), 96kHz sampling frequency at 24 bits.(However, two speakers, p280 and p315, had technical difficulties with the MKH 800 audio recordings.) All recordings were downsampled to 48 kHz and manually end-pointed after being converted to 16 bits. This corpus was created with HMM-based text-to-speech synthesis systems in mind, particularly speaker-adaptive HMM-based speech synthesis that use average voice models trained on numerous speakers as well as speaker adaption methods. This corpus is also suited for neural waveform modeling and DNN-based multi-speaker text-to-speech synthesis systems.

2.2.2 Preprocessing

The audio was preprocessed the same way as IEMOCAP, except for emotion label part, as there is no emotion label in VCTK.

2.3 Emotion Conditioned Speech Synthesis with IEMOCAP Dataset

Figure 1 shows the suggested model, which is based on a modified Tacotron2 [1–3] with an emotion embedding learning network, an auxiliary learning network, speaker embedding layer and using style loss.

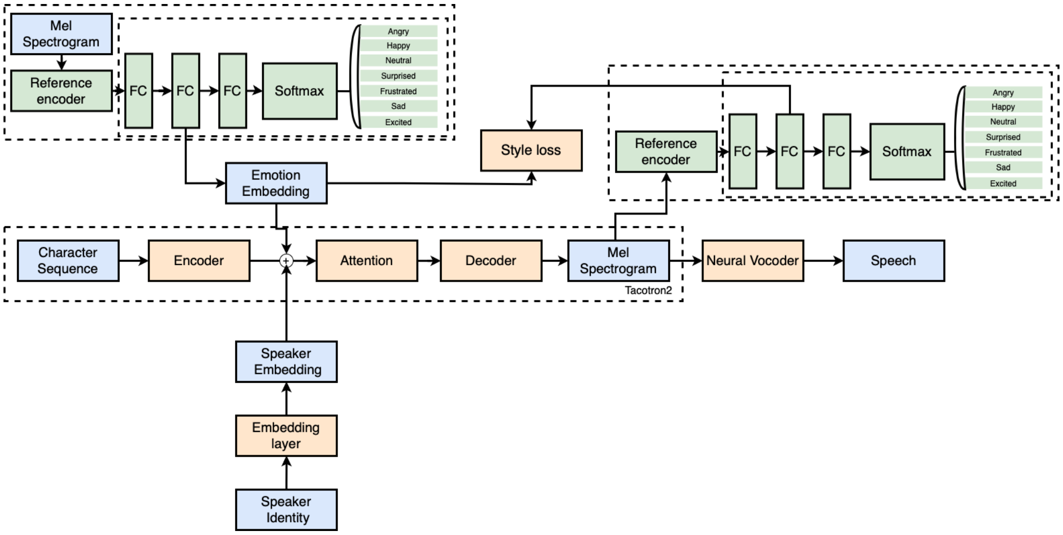


Figure 1: The architecture of the model

2.3.1 Reference encoder

The reference encoder takes a melspectrogram of length L_R and dimension d_R as input and returns an d_P embedding as described by Skerry-Ryan *et al.* [11]. The encoder network consists of six convolutional layers, a GRU layer followed by a 128 dimensional fully connected layer.

2.3.2 Emotion classifier

The emotion classifier takes the output of the reference encoder and outputs an emotion classification. Composed of three fully connected layers and a softmax layer the network tries to learn a better emotion embedding from the output of the reference encoder. The output of the second fully connected layer is used as the emotion embedding which we condition Tacotron2 on. During inference a melspectrogram of a desired emotion recording is used and the emotion embedding is multiplied by a scalar to control the strength of the emotion.

2.3.3 Tacotron2

Tacotron [12] is a sequence to sequence autoregressive text-to-speech network, which takes character sequence as input and outputs a melspectrogram. During training it doesn't need phoneme-level alignment, instead the alignment is learned through non-monotonic attention. It is trained with a MSE loss for the melspectrogram and the BCE loss for stop token prediction. Tacotron2 [3] is a slightly modified version.

2.3.4 Style Loss

The Gram matrix has recently been used to assess the mel-spectrogram of audio signals, with the goal of capturing local characteristics in the frequency-time domain. The Gram matrix is thought to be capable of representing low-level speech properties such as volume, stress, speed, pitch, and others that are closely linked to emotion displays. The suggested model's emotion embedding is a collection of CNN output sequences, which can be thought of as the mel-feature spectrogram's map. Each value of a feature map is derived from the convolution of a certain filter at a target region, and the essential is feature extraction and quantification, thus each value may be interpreted as the intensity of emotion-related characteristics. The objective is to synthesis speech with a certain target emotion category (for example, surprise) while managing the intensity of the emotion communicated to the target. The emotion style difference between the generated and reference speech is quantified utilizing style loss to accomplish this. The appropriate gram matrices G and I are constructed via inner-product given the feature map of reference and synthesized speech R and S:

$$G = R^T * R, and\, I = S^T * S$$

Basically, the more correlated the feature maps are the more style information is share between generated and reference. Since both R and S are originally vectors in our setting (activations from a fully connected layer), they first need to be reshaped into square matrices. When minimizing the loss L_{sty} between these Gram matrices we force the distribution of features to match between the emotion embeddings:

$$L_{sty} = \frac{1}{(2NM)^2} \sum (I - G)^2$$

where the N, is the number of rows and M, the number of columns of the matrix. The total loss is calculated as follows:

$$L_{total} = L_{tac} + L_{sty} + L_{cls-src} + L_{cls-tgt}$$

where L_{tac} is the Tacotron MSE loss, L_{sty} is the style loss, and$L_{cls-src}$ and $L_{cls-tgt}$ are both the classification (cross-entropy) loss of the encoder-classifier networks at encoding and decoding steps of the Tacotron.

2.4 Speaker Conditioned Speech Synthesis with VCTK Dataset

This time the problem we approach is different. Instead of transferring emotion to the generated speech, we transfer the style of a given speaker throughout a speaker embedding that aims to represent prosody characteristics of such speaker. The architecture, as seen in figures 3 and 2 is similar to the previous one used in the case of emotion conditioned speech synthesis, but much simpler now. We remove the encoder-classifier networks that generated and evaluated the emotional embeddings. Now Tacotron2 is only conditioned on some sort of a speaker embedding.

Three main approaches were considered to compare performance, they all use the Tacotron loss and only differ in the generation of the embedding:

1. In figure 2 the speaker embedding starts from a random initialization in a 512-d space and gets fully trained as the Tacotron trains.

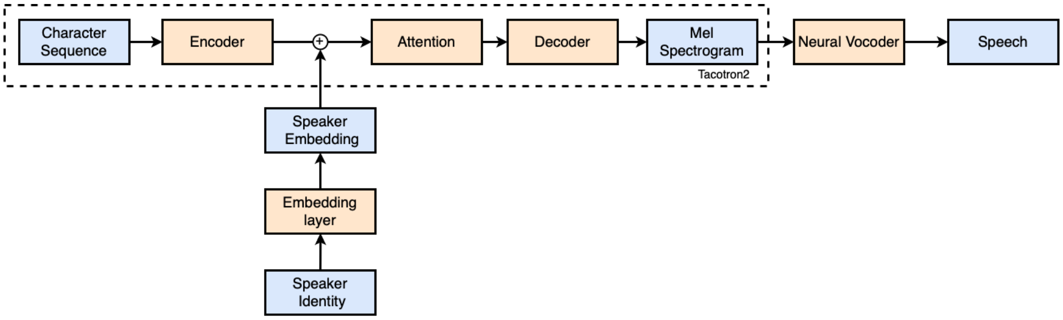


Figure 2: The architecture of the model with trainable embeddings

2. In figure 3, we use fixed embeddings generated by pre-trained speaker verification network [13]. The same approach is done in [14]. The advantage of this approach is that we can run inference also on speakers unseen during training.

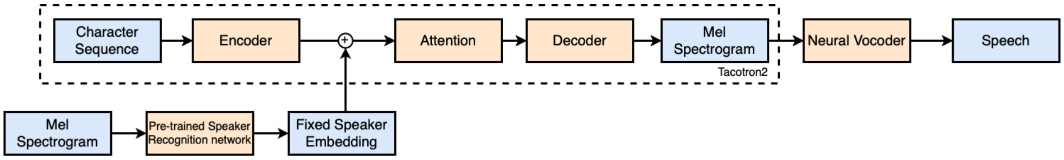


Figure 3: The architecture of the model with pretrained speaker verification network [13]

3. Finally, in figure 4 we use an encoder as the one described in section 2.3.1 and attention to learn weights for 10 random 256 dimensional tokens. This approach follows [15]. The training is unsupervised (doesn't require any style labels - speakers, emotions, accents, ...)

and the token weights therefore capture the the whole speaking style, not just one particular aspect.

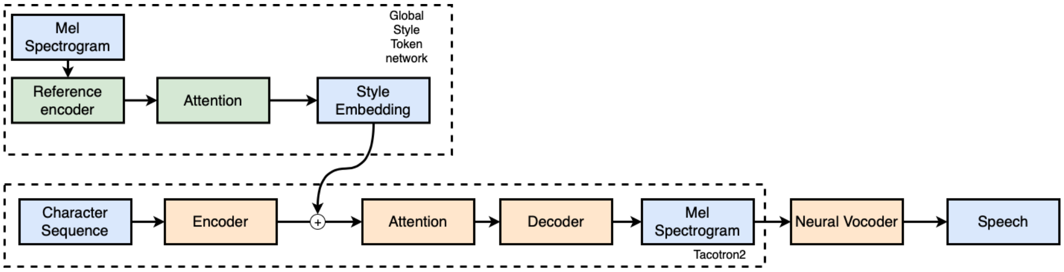


Figure 4: The architecture of the model with global style tokens.

3 Experiments

3.1 Emotional Multi-speaker Speech Synthesis on Tacotron2 with IEMOCAP Dataset

Aware of our different dataset we had to slightly diverge from the method in [1]. First of all we had to add the speaker embedding to account for multiple speakers. This was done through a simple embedding layer conditioned on the speaker's identity. Secondly, we didn't change the Tacotron2 architecture as they did in [1] because of preserving simplicity and because we wanted to start training from pretrained weights given our small dataset. For the same reason the speaker embedding was added to every encoder output instead of concatenation. We started training multi-speaker Tacotron2 without the emotion embedding, however, none of the above was enough for Tacotron2 to start forming the alignments between text and speech. Also, because of our hardware limitations, we only train with batch size 6. Furthermore, when examining the data with a pretrained speaker verification network [13] and also examining our trained embeddings, in 5 and 6 we can see that in both the speakers from the same conversations are clustered together. This suggests that the sentences are not perfectly separated and that the overlapping conversations influence the model. Therefore, we concluded that the lack of data (6.4h) and its quality (half was spontaneous conversation) in this dataset are the cause of the model failure.

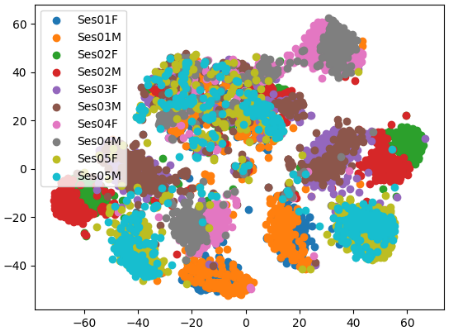


Figure 5: Embeddings from pretrained speaker verification network on the train data (tSNE projection)

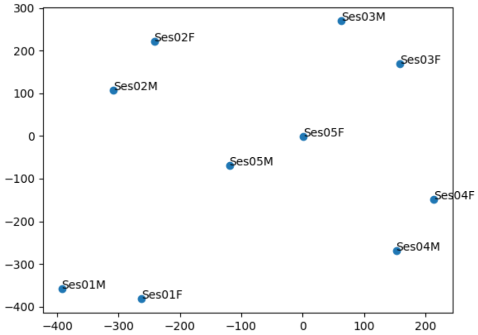


Figure 6: Embeddings learned by Tacotron2 12000 update steps (tSNE projection)

3.2 Multi-speaker Speech Synthesis on Tacotron2 with VCTK Dataset

Realizing that we lacked the amount of labeled data required to train Tacotron2 we decided to ignore the emotional speech part and just focus on training Tacotron2 on multiple speakers. We therefore switched from the IEMOCAP dataset to the VCTK dataset and switched to only focusing on producing a good multi-speaker model. In figure 7 we can observe that in the VCTK data the speakers can be separated much more easily than in IEMOCAP. Furthermore, in the PCA projection in Figure 8 we see that the speakers can also be separated by gender and accents, even though with several outilers. The experiments with this dataset consist of comparing the 3 methods for speaker embedding presented in the previous section.

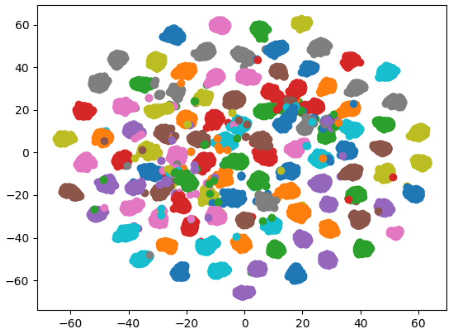


Figure 7: Embeddings from pretrained speaker verification network on the train data (tSNE projection) (some speakers share the same colour)

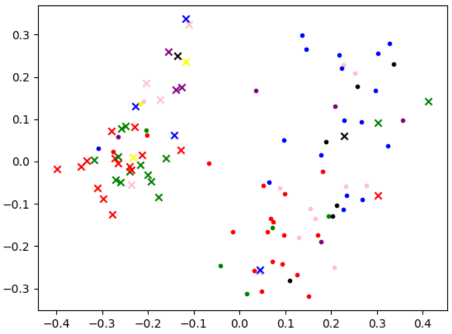
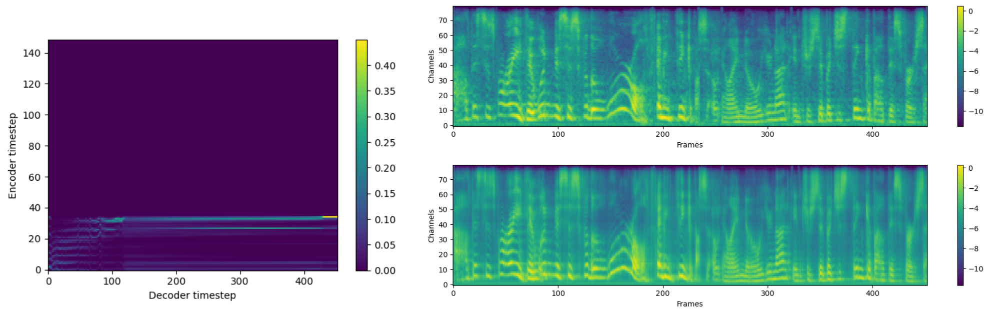


Figure 8: Averaged embeddings from pretrained speaker verification network on the train data (PCA projection) (Different colors show different accents, cross are male, circle female voices)

4 Results

4.1 Emotional Multi-speaker Speech Synthesis on Tacotron2 with IEMOCAP Dataset

In figure 9 we can see a validation sequence example. The ground truth and predicted mel-spectrograms look very similar, however this is not indicative of the success of the training, because teacher forcing is used here. It is important to see whether the attention alignment has been formed, which would resemble a diagonal line.



(a) Attention alignment

(b) Ground Truth mel-spectrogram (top), predicted mel-spectrogram (bottom)

Figure 9: Validation sequence example during training (12000 update steps)

In figure 10 we can see an example of an inference of the sentence "I speak your language", which is present in the training set for speaker Ses01M. Left is melspectrogram before post-net, center is after post-net and right is the attention alignment. We see that the model here completely fails (no teacher forcing here).

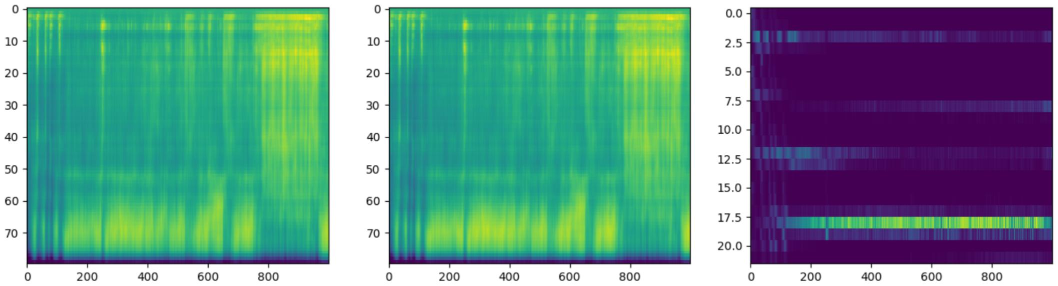


Figure 10: Inference sequence example after 12000 update steps from warm-up model

4.2 Multi-speaker Speech Synthesis on Tacotron2 with VCTK Dataset

Training with the VCTK dataset was successful and the attention alignment started forming very early (already after 2000 update steps) as seen in figure 11.

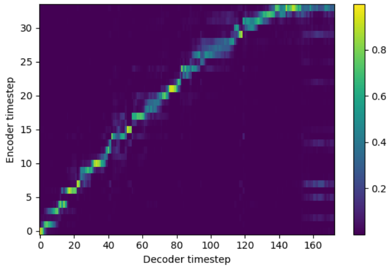


Figure 11: Validation sequence example during training (2000 update steps)

In general female voices seem better, maybe because the pre-trained model was trained on a female voice. The synthesized speech is usually evaluated qualitatively by MOS (Mean opinion score) and compared with different methods, but due to time constraints we did not perform this evaluation. However, by listening to our generated speech we noticed the differences between our 3 approaches:

1. **Trainable embedding** - The best quality, but sometimes the stop token is not predicted correctly (long silence at the end).
2. **Fixed embedding** - Sometimes big pauses between words.
3. **Global style tokens** - Has the most artifacts (repeating words, robotic sounds, noise)

Our **Fixed embedding** model can also successfully condition on unseen speakers. The quality is understandably lower than on seen speakers, however the speech is still intelligible and the difference between male and female speakers is also clearly noticeable.

5 Discussion and Conclusions

Based on our experiments it is clear that Tacotron2 requires a lot of good quality data to train properly, literature [5; 8] blames this on the non-monotonic attention alignment mechanism. This might also cause skipping or repeating words. There are some text-to-speech models e.g. [8] that circumvent these problems, but as that would require us to rework large parts of the project we settled with switching datasets instead. This forced us to abandon our emotion conditioning intention. Comparing the three approaches to speaker conditioning, the Trainable embedding provides the most flexibility and information and reaches the highest quality. Meanwhile, the Fixed embedding doesn't lack much in quality and allows to generalize on unseen speakers. The GST approach probably requires much more tuning, but allows training without any conditioning labels. For future work it might be interesting to explore whether the model is able to learn to condition on different English accents (independently of the speaker's identity) and further explore the generalization on unseen speakers.

References

[1] T. Li, S. Yang, L. Xue, and L. Xie, “Controllable emotion transfer for end-to-end speech synthesis,” in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1–5, 2021.

[2] NVIDIA, “Tacotron2.” https://github.com/NVIDIA/tacotron2, 2018.

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” 2017.

[4] NVIDIA, “Waveglow.” https://github.com/NVIDIA/waveglow, 2018.

[5] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, “A survey on neural speech synthesis,” 2021.

[6] J. Kim, S. Kim, J. Kong, and S. Yoon, “Glow-tts: A generative flow for text-to-speech via monotonic alignment search,” 2020.

[7] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fastspeech: Fast, robust and controllable text to speech,” 2019.

[8] S. Mehta, E. Szekely, J. Beskow, and G. E. Henter, “Neural HMMS are all you need (for high-quality attention-free TTS),” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, may 2022.

[9] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, “Neural speech synthesis with transformer network,” 2018.

[10] J. Yamagishi, C. Veaux, and K. MacDonald, “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” The Centre for Speech Technology Research (CSTR), 2019.

[11] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” 2018.

[12] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” 2017.

[13] resemble ai, “Resemblyzer.” https://github.com/resemble-ai/Resemblyzer, 2019.

[14] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” 2018.

[15] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” 2018.

[16] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *Computer Vision – ECCV 2016* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 694–711, Springer International Publishing, 2016.

[17] L. Qin, Z.-H. Ling, Y.-J. Wu, B.-F. Zhang, and R.-H. Wang, “Hmm-based emotional speech synthesis using average emotion model,” in *Chinese Spoken Language Processing* (Q. Huo, B. Ma, E.-S. Chng, and H. Li, eds.), (Berlin, Heidelberg), pp. 233–240, Springer Berlin Heidelberg, 2006.

[18] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, “Deep voice 2: Multi-speaker neural text-to-speech,” 2017.