Improving Multi-Speaker Tacotron with Speaker Gating Mechanisms

Wei Zhao[1,2], Li Xu[1,2], Ting He[1,2]
[1]College of Electrical Engineering, Zhejiang University, Hangzhou, Zhejiang 310027, China
[2]Zhejiang University Robotics Institute, Yuyao, Zhejiang 315400, China
E-mail: weizhao_ee@zju.edu.cn

Presenter Name Card
Meeting Code: 292 903 58463

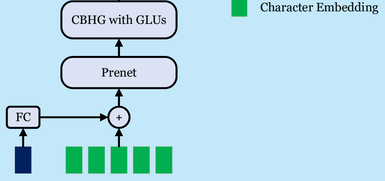
Abstract

In this paper, we present two speaker gating mechanisms for multi-speaker Tacotron, a popular end-to-end text-to-speech (TTS) neural system, to improve the performance of generating multiple voices. With our presented mechanisms, the model can work better in both generalization and accuracy. As a starting point, we introduce the original multi-speaker Tacotron as a baseline model because of its excellent performance and straightforward structure. Employing gated linear units (GLUs), two different speaker gating mechanisms are then proposed for this model. Extensive experiments on VCTK dataset are conducted to demonstrate the validity of our methods. Conclusively, we find that it is promising to incorporate the speaker identity information by using the proposed speaker gating mechanisms.

Gating Mechanism A

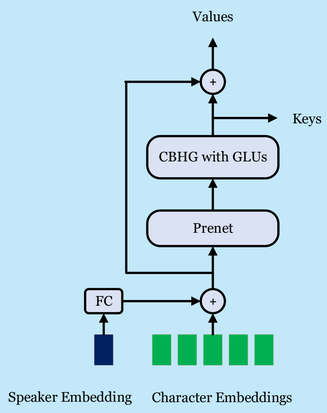
We hold a view that outputs of the encoder in multi-speaker Tacotron should be various for multiple speakers to promote the performance of the system. For that reason, GLUs is utilised to replace the bank of 1-D convolution units in the CBHG module.

The modified structure of the text encoder is depicted as follow:



Gating Mechanism B

For a further study, a modified structure of the encoder with key-value output pairs is also demonstrated as follows.



where h_j is the j-th state of attention RNN, $a_{i,j}$ is the attention alignments, and c_j is the context vector. e_i represents the keys and x_i represents the point information about a specific input element.

$$a_{i,j} = \text{Attend}\,(h_j, e_i) \tag{1}$$

$$c_j = \sum_{i=1}^{n} a_{i,j}\,(x_i + e_i) \tag{2}$$

$$y_j = \text{Generate}\,(h_j, c_j) \tag{3}$$

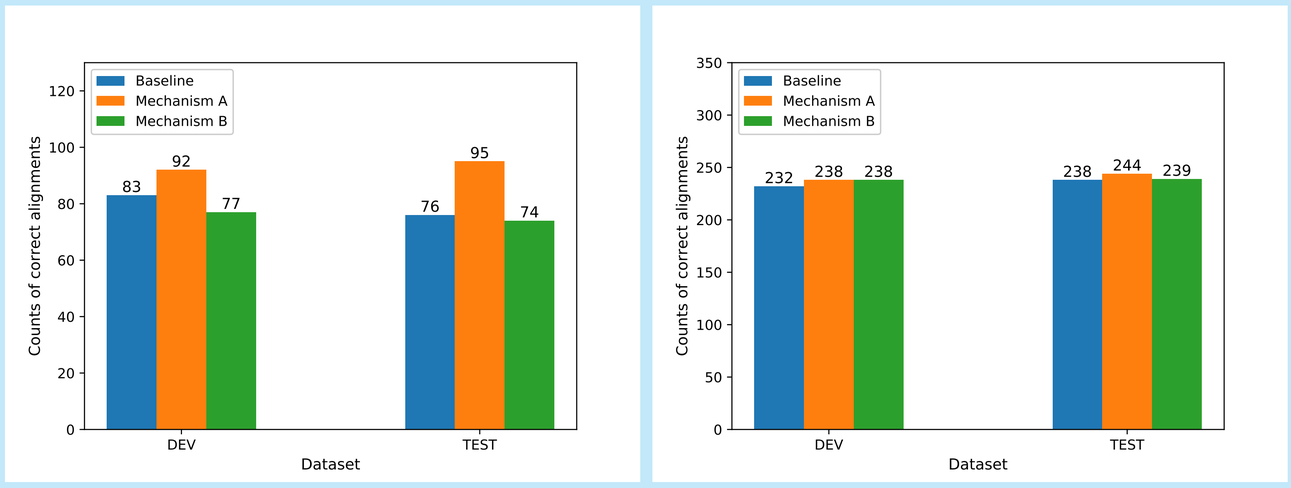
Multi-Speaker Tacotron

The original multi-speaker Tacotron consists of an encoder, an attention-based decoder, and a post-processing net. The encoder takes character embeddings as inputs and extracts sequential representations. The decoder produces target Mel-scale spectrograms which are conditioned on the concatenation of encoder outputs and speaker embeddings at each time step with an attention mechanism. Specifically, the speaker embedding is broadcast over encoder outputs for the concatenation operation. The post-processing net is adopted to generate linear spectrograms with appropriate dimension. Finally, via feeding linear spectrograms to the Griffin-Lim vocoder, the system outputs corresponding speeches.

Figures

we evaluated the generalization capability of models by counting attention failures which were meaningless. The fewer failures the model does, the better performance it has. For the VCTK-24 subset, we selected 5 transcriptions for each speaker and for the VCTK-84 subset, we selected 3 transcriptions for each speaker. Figure 1 gives a comparison of the results of the baseline model and our modified models.

Figure 1: The statistics of correct alignments on VCTK-24 and VCTK-84



Tables

Besides the generalization, accuracy is also very important, especially for multi-speaker system. We measure the accuracy via a speaker identification system, based on Gaussian Mixture Model (GMM) and trained on both VCTK-24 and VCTK-84 dataset. We toke the Top-1 identification accuracy to quantify the corresponding performance of the models.

Table 1: Multi-Speaker Identification Top-1 Accuracy (%)

Model	DEV-24	TEST-24	DEV-84	TEST-84
Ground Truth	97.56	97.56	94.48	94.48
Baseline	87.80	89.74	55.92	49.66
Mechanism A	90.24	**94.87**	56.58	48.97
Mechanism B	**92.68**	92.31	**56.58**	**53.79**

In order to assess the influence of our modifications on speech quality, we employed the Mel Cepstral Distortion (MCD) scores.

Table 2: Multi-speaker MCD scores (Mean; lower is better)

Model	DEV-24	DEV-84
Baseline	24.67±3.33	21.33±1.76
Mechanism A	**22.61±2.91**	21.31±1.93
Mechanism B	23.11±2.69	**20.86±1.87**

Conclusion

Although there already have been many excellent end-to-end works to synthesize multiple voices, most of them do not focus on optimizing and analysing the approach of incorporating speaker embeddings. In this work, we propose the speaker gating mechanisms to improve the multi-speaker Tacotron in both generalization and accuracy. The experimental results show that model with mechanism A has a better performance in generalization, especially for a smaller dataset. At the same time, the model with mechanism B performs more predominately in accuracy and quality, and is more robust over datasets with different size.