TEACHER-STUDENT TRAINING FOR ROBUST TACOTRON-BASED TTS

Rui Liu[1,2] , *Berrak Sisman*[2,4], *Jingdong Li*[1], *Feilong Bao*[*1], *Guanglai Gao*[1], *Haizhou Li*[2,3]

[1] Inner Mongolia University, China [2] National University of Singapore, Singapore
[3] University of Bremen, Germany [4] Singapore University of Technology and Design (SUTD)
{r.liu,berraksisman}@u.nus.edu, haizhou.li@nus.edu.sg

ABSTRACT

While neural end-to-end text-to-speech (TTS) is superior to conventional statistical methods in many ways, the exposure bias problem in the autoregressive models remains an issue to be resolved. The exposure bias problem arises from the mismatch between the training and inference process, that results in unpredictable performance for out-of-domain test data at run-time. To overcome this, we propose a teacher-student training scheme for Tacotron-based TTS by introducing a distillation loss function in addition to the feature loss function. We first train a Tacotron2-based TTS model by always providing natural speech frames to the decoder, that serves as a teacher model. We then train another Tacotron2-based model as a student model, of which the decoder takes the predicted speech frames as input, similar to how the decoder works during run-time inference. With the distillation loss, the student model learns the output probabilities from the teacher model, that is called knowledge distillation. Experiments show that our proposed training scheme consistently improves the voice quality for out-of-domain test data both in Chinese and English systems.

Index Terms— Tacotron, Knowledge Distillation, TTS

1. INTRODUCTION

With the advent of deep learning, end-to-end TTS has shown many advantages over the conventional TTS techniques [1–3]. For example, Tacotron-based approaches [4–7] with an encoder-decoder architecture and attention mechanism have shown to achieve remarkable performance. In these techniques, the key idea is to integrate the conventional TTS pipeline into a unified network and learn the mapping directly from the <text, wav> pair [7–10]. Furthermore, together with a neural vocoder [5, 11–15], natural-sounding human-like speech can be generated.

However, neural end-to-end TTS is still far from perfect. A typical neural TTS system suffers from the exposure bias problem [16, 17] in the autoregressive model [18] that is used by the decoder module. Specifically, in training stage, the decoder generates a frame using its previous frames of natural speech as input, that is called *teacher forcing mode*. However, in inference stage, the decoder predicts a frame using its previously predicted frames as input, that is also called *free running mode*. There exists a mismatch between the natural speech frames and the predicted frames especially for out-of-domain test data, that leads to unpredictable outcomes, such as skipping, repeating words, incomplete synthesis and inappropriate prosody phrase breaks [9, 19–21].

The techniques to improve in-domain performance of neural TTS frameworks include attention mechanism [22] and scheduled sampling [23, 24]. The use of scheduled sampling comes with negative effects that include misalignment between the natural speech frames and the predicted frames due to the fact that the temporal dependency of the acoustic sequence is disrupted. The techniques to improve out-of-domain performance include the GAN-based TTS framework [25] that introduces both real and generated data sequences in discriminator training, and more recently, stepwise monotonic attention for neural TTS [9].

In this paper, we propose a novel training scheme, the teacher-student training scheme, for neural end-to-end TTS framework, that performs remarkably well for out-of-domain inference. In this scheme, a teacher model learns the text-speech mapping from training data in teacher forcing mode, while a student model learns from both the probability distribution of the teacher model and the same training data for teacher model in free running mode. The process of student learning from teacher model is called knowledge distillation, and its learning objective is called distillation loss.

The main contributions of this paper are summarized as follows: 1) we propose a compact method for end-to-end TTS model; and 2) we propose a teacher-student training scheme for Tacotron-based TTS model. To our best knowledge, this is the first implementation of teacher-student training scheme for Tacotron2 based TTS framework. The proposed training scheme is validated with out-of-domain test data in Chinese and English TTS systems.

This paper is organized as follows: In Section 2, we re-visit the Tacotron2-based TTS framework that serves as a baseline reference. In Section 3, we study the proposed teacher-student training scheme. In Section 4, we report the evaluation results. We conclude the paper in Section 5.

Speech samples: https://ttslr.github.io/ICASSP2020/

2. TACOTRON2 BASED TTS

In this paper, we adopt Tacotron2 [5] with scheduled sampling in the training stage, as a reference baseline. For rapid turnaround, we use Griffin-Lim [26] waveform reconstruction instead of WaveNet vocoder in this study. We note that the selection of waveform generation technique will not affect our judgment of the effectiveness of the proposed training scheme.

We illustrate the overall architecture of the *reference baseline* in Figure 1, that includes encoder, attention-based decoder and Griffin-Lim algorithm. The encoder consists of two components, a CNN [27, 28] based module that has 3 convolution layers, and a LSTM [29, 30] based module that has a bidirectional LSTM layer. The decoder consists of four components: a 2-layer pre-net, 2 LSTM layers, a linear projection layer and a 5-convolution-layer post-net. The decoder is a standard autoregressive recurrent neural network that generates the mel-spectrogram features and stop tokens frame by frame.

During training, the decoder generates a frame in the scheduled sampling mode. However, at run-time inference, the decoder performs in free running mode to predict the future frames. Such trained decoder experiences the mismatch between the natural speech frames and the predicted speech frames, and the adverse effect of scheduled sampling on the temporal dependency of natural acoustic sequence. To address the above issues during training, we study a teacher-student training scheme in Section 3.

3. TEACHER-STUDENT TRAINING FOR TACOTRON2 BASED TTS

In this section, we discuss in detail the teacher model, the student model, and the teacher-student training scheme. While both the teacher model and the student model have identical network architecture as the *reference baseline* , they adopt different decoding strategies as illustrated in Figure 2.

In practice, we first train a standard Tacotron2 teacher model for an end-to-end TTS system under the *teacher forcing mode*, that is regarded as the teacher model. As the teacher model learns under the *teacher forcing mode*, it is expected to represent the true distribution of the natural speech data. We then train another Tacotron2 student model under the *free running mode*. The student model is trained by learning from both ground-truth sequence and the hidden states of the teacher model simultaneously. By learning from the hidden states of the teacher model via knowledge distillation, the student model learns the true distribution of the natural speech data effectively. As the student model is trained under the *free running mode* by using the predicted speech frames as the input of the decoder, it is expected to accustom itself to the run-time inference condition.

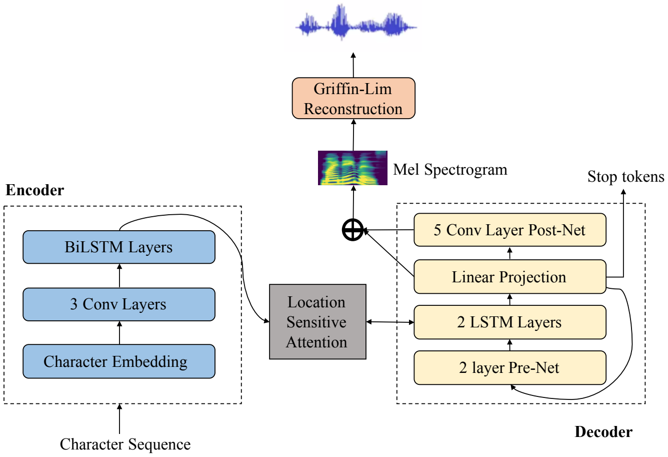

Fig. 1. Block diagram of Tacotron2-based reference baseline that has three modules, encoder, attention-based decoder, and Griffin-Lim reconstruction algorithm.

3.1. Teacher Model

For the decoder in teacher model, we implement the *teacher forcing mode* that predicts a speech frame by taking the previous natural speech frames in the sequence as the input.

Given a input character sequence $x = (x_1, x_2, ..., x_T)$ and its target mel-spectrogram features $y = (y_1, y_2, ..., y_{T'})$, let $P(y|x, \theta)$ is the teacher model of which θ is the model parameters. Teacher model with teacher forcing mode takes the previous frames $y_1, ..., y_{t-1}$ from the target natural speech as input to predict the feature frame y_t at time step t, as formulated next,

$$P(\hat{y}|x, \theta) = \prod_{t=1}^{T'} P(\hat{y}_t | y_{<t}, x, \theta) \tag{1}$$

where $\hat{y}$ is the predicted value and y is from the target natural speech.

With such decoding mode, the teacher model is expected to learn the true probability distribution from natural speech data, that would be very informative for the student model.

3.2. Student Model

The student model has the same network architecture as the teacher model, except that it has a completely different decoding mode: *free running mode*. In this mode, the decoder predicts a speech frame by taking the previous predicted speech frames in the sequence as the input. The decoding process of the student model is defined as:

$$P(\hat{y}|x, \theta) = \prod_{t=1}^{T'} P(\hat{y}_t | \hat{y}_{<t}, x, \theta) \tag{2}$$

where $\hat{y}$ is the predicted value.

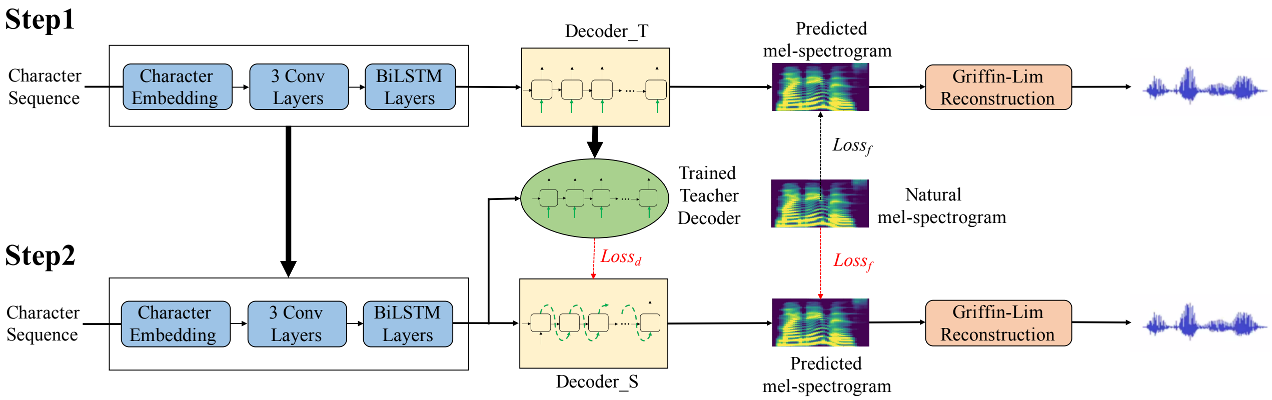

Fig. 2. Illustration of the proposed teacher-student training scheme for Tacotron2-based TTS in 2 steps: Step 1, pre-train a teacher model, that includes teacher encoder and teacher decoder ("Decoder_T"); Step 2, use the trained teacher encoder and teacher decoder to train the student decoder ("Decoder_S") by applying the proposed knowledge distillation approach.

3.3. Knowledge Distillation

Typically, knowledge distillation is a process where a small model is trained to mimic a pre-trained, larger model [31]. In this paper, we borrow the idea of knowledge distillation in the implementation of the teacher-student training scheme.

The idea is to use a teacher model, that has been trained under the *teacher forcing mode*, to guide the training of the student model, that runs under free running mode. As the teacher model is trained using natural speech frames as the input of decoder, we expect the output probability distribution of the teacher model to reflect the true distribution of the natural speech data. The student model is trained under the *free running mode*. Therefore, it is closer to the actual inference condition. At the same time, the hidden states of the student model are optimized to be close to those of the teacher model by way of knowledge distillation. As can be seen in Figure 2, we define one objective function for the teacher model, the feature loss. We devise two objective functions for the student model, one for the feature loss that is the same as in the teacher model, and another for the knowledge distillation, or distillation loss.

We formulate the entire process next. The encoder takes the input character sequence $x = (x_1, x_2, ..., x_T)$ from the given text and converts the one-hot vector to continuous high-level features representation h:

$$h_t = \text{Encoder}(h_{t-1}, x_t) \tag{3}$$

The teacher decoder, *Decoder_T* outputs a hidden state s_t at each step t:

$$s_t = \text{Decoder\_T}(s_{t-1}, \hat{y}_{t-1}, \sigma(h_t)) \tag{4}$$

where $\sigma()$ represents a function to calculate the context vector by using location-sensitive attention mechanism.

Similarly, the student decoder *Decoder_S* processes the same input sequence and generates the hidden state $\hat{s}_t$ at each step t at the same time:

$$\hat{s}_t = \text{Decoder\_S}(\hat{s}_{t-1}, \hat{y}_{t-1}, \sigma(h_t)) \tag{5}$$

In both the teacher model and the student model, the feature loss function $Loss_f$ ensures that the generated speech is close the the target speech,

$$Loss_f = \sum_{t=1}^{T'} L_2(\hat{y}_t, y_t) \tag{6}$$

In the student model, to minimize the discrepancy between the hidden states s and $\hat{s}$ of the teacher model and the student model, we introduce the distillation loss $Loss_d$,

$$Loss_d = \frac{1}{T}\sum_{t=1}^{T} |s - \hat{s}|^2 \tag{7}$$

Then the total loss function for the student model is therefore,

$$Loss_{total} = Loss_f + \lambda \cdot Loss_d \tag{8}$$

where λ is a trade-off parameter for the two loss terms.

With knowledge distillation, the proposed 2-step teacher-student training scheme allows for a more compact End-to-End network than others such as generative adversarial network [25]. The teacher model is trained with the objective function $Loss_f$ under the teacher forcing mode, while the student model is trained with a combination of two loss functions $Loss_{total}$ under the free running mode.

4. EXPERIMENTS

We develop two systems on Chinese (12 hours of Data Baker [1]) and English (LJSpeech [2]) corpora separately. To verify the effectiveness of knowledge distillation, denoted as *Tacotron2-KD*, we choose 2 baseline frameworks: 1) Tacotron2 with scheduled sampling, denoted as *Tacotron2-SS*, and 2) Tacotron2 with free running mode, denoted as *Tacotron2-FR*. In all experiments, we use Griffin-Lim algorithm [26] for waveform generation for rapid turn-around.

[1] https://www.data-baker.com/open_source.html
[2] https://keithito.com/LJ-Speech-Dataset/

Framework	Language	MOS	WER
Tacotron2-SS	en	3.21	23.82%
	cn	3.18	9.44%
Tacotron2-KD	en	**3.93**	**2.17%**
	cn	**3.94**	**0.67%**

Table 1. Comparison of mean opinion scores (MOS) and Word Error Rate (WER%) between Tacotron2-SS and the proposed Tacotron2-KD.

4.1. Experimental Setup

For Chinese experiments, the encoder takes pinyin sequence with tones as input and generates an 160-channel Mel spectrum, two frames at a time, as output. For English experiments, the encoder takes the character sequence as input and generates an 80-channel Mel spectrum, two frames at a time, as output. The two type of encoder inputs are collectively referred to as *character* in this paper. For both systems, we use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 10^{-3} exponentially decaying to 10^{-5} starting after 50k iterations. We also apply L_2 regularization with weight 10^{-6}. Hyper-parameter λ in Equation 8 is set as 1.0 and all the models are trained with a batch size of 32. In teacher-student model training, we adopt the teacher model trained with 150k steps as the teacher decoder "Decoder_T", and train the student decoder "Decoder_S" for 150k steps with the proposed knowledge distillation method.

4.2. Subjective Evaluation

We conduct experiments with out-of-domain test data for naturalness and robustness evaluation. For Chinese, we select 500 test samples from the Blizzard Challenge 2019 Chinese dataset [32]. For English, we select 50 test samples from FastSpeech [20], which are particularly hard for TTS system. In addition to the 50 test samples, that are single letters, spellings, repeated numbers, we also include 30 long sentences, each having 128 characters on average. 20 English speakers and 15 Chinese speakers participated in the listening tests. Each subject listens to 80 converted utterances of his/her native language.

4.2.1. Naturalness Evaluation

We first evaluate the sound quality of the synthesized speech with mean opinion score (MOS) among Tacotron2-SS, Tacotron2-FR and the proposed Tacotron2-KD, that is reported in Table 1. The listeners rate the quality on a 5-point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. It is observed that the proposed Tacotron2-KD clearly outperforms the baseline Tacotron2-SS for both English and Chinese data. As we observe that Tacotron2-FR achieves MOS of 1.33 for English and 2.32 for Chinese, that is significantly lower than those of Tacotron2-SS, we exclude Tacotron2-FR in AB preference test.

The AB preference test is reported in Figures 3 and 4, to compare Tacotron2-KD and Tacotron2-SS, in terms of voice quality. It is observed that Tacotron2-KD outperforms Tacotron2-SS consistently for both English and Chinese data.

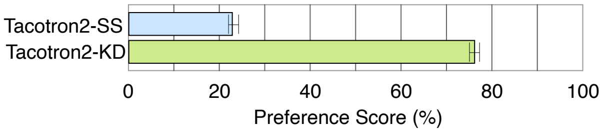

Fig. 3. The preference test between Tacotron2-KD and Tacotron2-SS on English data, with 95% confidence interval.

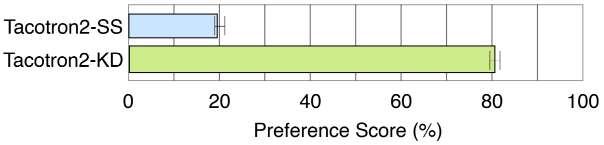

Fig. 4. The preference test between Tacotron2-KD and Tacotron2-SS on Chinese data, with 95% confidence interval.

4.2.2. Robustness Evaluation

We further conduct experiments to evaluate the robustness of synthesized speech for Tacotron2-SS and the proposed Tacotron2-KD, as reported in Table 1. We measure the robustness by Word Error Rate (WER %), that reports the sum of repeats (insertions) and skips (deletions) over the total number of characters in the listening tests [25]. Repeats and skips represent the two types of errors that Tacotron2 faces. It is shown that Tacotron2-KD effectively reduces the errors by 8.77% and 21.65% over the Tacotron2-SS baseline.

A detailed analysis finds that Tacotron2-SS generates 528 skips and 9 repeats for Chinese data, and 251 skips and 12 repeats for English data, while Tacotron2-KD generates only 24 skips for English data and 38 skips for Chinese data. We don't observe any repeats from the Tacotron2-KD outputs, that we think is remarkable.

5. CONCLUSION

We have studied a training scheme for Tacotron2 to perform high-quality speech synthesis for out-of-domain text, that overcomes the exposure bias problem. We implement the teacher-student training scheme through a knowledge distillation objective function. We have conducted a series of experiments on both Chinese and English to evaluate the naturalness and robustness. The proposed Tacotron2-KD framework consistently outperforms the baseline systems in both languages.

In addition to the naturalness and robustness improvement, we also discover that Tacotron2-KD delivers improved prosody renderings especially. We will report the prosody analysis of Tacotron2-KD system in the future.

6. ACKNOWLEDGEMENTS

This research was supported by Human-Robot Interaction Phase 1 (Grant No. 192 25 00054) project, National Research Foundation, Prime Minister's Office, Singapore under the National Robotics Programme, and by the National Natural Science Foundation of China (No.61563040, No.61773224), Natural Science Foundation of Inner Mongolian (No.2018MS06006, No.2016ZD06) and China Scholarship Council.

7. REFERENCES

[1] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[2] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 ieee international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7962–7966.

[3] Rui Liu, Feilong Bao, Guanglai Gao, and Yonghe Wang, "Mongolian text-to-speech system based on deep neural network," in *National Conference on Man-Machine Speech Communication*. Springer, 2017, pp. 99–108.

[4] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: A fully end-to-end text-to-speech synthesis model," in *INTERSPEECH*, 2017, pp. 4006–4010.

[5] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[6] Rui Liu, Berrak Sisman, Feilong Bao, Guanglai Gao, and Haizhou Li, "Wavetts: Tacotron-based tts with joint time-frequency domain loss," in *arXiv 2002.00417*, 2020.

[7] Younggun Lee and Taesu Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *ICASSP 2019-IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5911–5915.

[8] Yu-An Chung, Yuxuan Wang, Wei-Ning Hsu, Yu Zhang, and RJ Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6940–6944.

[9] Mutian He, Yan Deng, and Lei He, "Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS," in *Proc. Interspeech 2019*, 2019, pp. 1293–1297.

[10] Hieu-Thi Luong, Xin Wang, Junichi Yamagishi, and Nobuyuki Nishizawa, "Training Multi-Speaker Neural Text-to-Speech Systems Using Speaker-Imbalanced Speech Corpora," in *Proc. Interspeech 2019*, 2019, pp. 1303–1307.

[11] Tomoki Hayashi, Akira Tamamori, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "An investigation of multi-speaker training for wavenet vocoder," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 712–718.

[12] Takuma Okamoto, Tomoki Toda, Yoshinori Shiga, and Hisashi Kawai, "Real-Time Neural Text-to-Speech with Sequence-to-Sequence Acoustic Model and WaveGlow or Single Gaussian WaveRNN Vocoders," in *Proc. Interspeech 2019*, 2019, pp. 1308–1312.

[13] Berrak Sisman, Mingyang Zhang, and Haizhou Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder," in *Proc. Interspeech 2018*, 2018, pp. 1978–1982.

[14] Berrak Sisman, Mingyang Zhang, and Haizhou Li, "Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2019.

[15] Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura, "Adaptive wavenet vocoder for residual compensation in gan-based voice conversion," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 282–289.

[16] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba, "Sequence level training with recurrent neural networks," in *Proc. ICLR 2016*.

[17] Florian Schmidt, "Generalization in generation: A closer look at exposure bias," *arXiv preprint arXiv:1910.00292*, 2019.

[18] Biing-Hwang Juang and Lawrence Rabiner, "Mixture autoregressive hidden markov models for speech signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 6, pp. 1404–1413, 1985.

[19] Rui Liu, Feilong Bao, Guanglai Gao, Hui Zhang, and Yonghe Wang, "Improving mongolian phrase break prediction by using syllable and morphological embeddings with bilstm model.," in *Interspeech*, 2018, pp. 57–61.

[20] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS 2019*, 2019.

[21] Xiaolian Zhu, Yuchao Zhang, Shan Yang, Liumeng Xue, and Lei Xie, "Pre-alignment guided attention for improving training efficiency and model stability in end-to-end speech synthesis," *IEEE Access*, vol. 7, pp. 65955–65964, 2019.

[22] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4784–4788.

[23] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 1171–1179.

[24] Ferenc Huszár, "How (not) to train your generative model: Scheduled sampling, likelihood, adversary?," *arXiv preprint arXiv:1511.05101*, 2015.

[25] Haohan Guo, Frank K. Soong, Lei He, and Lei Xie, "A New GAN-Based End-to-End TTS Training Algorithm," in *Proc. Interspeech 2019*, 2019, pp. 1288–1292.

[26] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[28] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham, "Learning attribute representations with localization for flexible fashion search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7708–7717.

[29] Kenan Emir Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf Kassim, "Semantically consistent hierarchical text to fashion image synthesis with an enhanced-attentional generative adversarial network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

[30] Mingyang Zhang, Berrak Sisman, Sai Sirisha Rallabandi, Haizhou Li, and Li Zhao, "Error reduction network for dblstm-based voice conversion," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 823–828.

[31] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4133–4141.

[32] Rui Liu, Jingdong Li, Feilong Bao, and Guanglai Gao, "The imu speech synthesis entry for blizzard challenge 2019," in *Proceedings of Blizzard_Challenge 2019*, 2019.