



Speaker consistency loss and step-wise optimization for semi-supervised joint training of TTS and ASR using unpaired text data

Naoki Makishima, Satoshi Suzuki, Atsushi Ando, Ryo Masumura

NTT Corporation, Japan

naoki.makishima.fx@hco.ntt.co.jp

Abstract

In this paper, we investigate the semi-supervised joint training of text to speech (TTS) and automatic speech recognition (ASR), where a small amount of paired data and a large amount of unpaired text data are available. Conventional studies form a cycle called the TTS-ASR pipeline, where the multi-speaker TTS model synthesizes speech from text with a reference speech and the ASR model reconstructs the text from the synthesized speech, after which both models are trained with a cycle-consistency loss. However, the synthesized speech does not reflect the speaker characteristics of the reference speech and the synthesized speech becomes overly easy for the ASR model to recognize after training. This not only decreases the TTS model quality but also limits the ASR model improvement. To solve this problem, we propose improving the cycle-consistency-based training with a speaker consistency loss and step-wise optimization. The speaker consistency loss brings the speaker characteristics of the synthesized speech closer to that of the reference speech. In the step-wise optimization, we first freeze the parameter of the TTS model before both models are trained to avoid over-adaptation of the TTS model to the ASR model. Experimental results demonstrate the efficacy of the proposed method.

Index Terms: Speech recognition, speech synthesis, self-supervision, semi-supervised learning

1. Introduction

Text to speech (TTS) and automatic speech recognition (ASR) perform inverse tasks of each other. Although these two models were developed independently, several recent studies have investigated learning both the TTS model and ASR model simultaneously [1–6]. One of the application in these studies is semi-supervised joint training, which creates models with a small amount of paired (speech and text) data and a large amount of unpaired text data. Compared to other semi-supervised training methods of TTS [7–9] and ASR [10–12], the advantage of semi-supervised joint training is that knowledge obtained by one model spreads to the other model, which is closely related to the speech chain mechanism of human communication [13]. Thus, semi-supervised joint training of TTS and ASR is a promising approach for applications that learn while speaking and listening like a human. In this paper, we focus on semi-supervised joint training utilizing paired data and unpaired text data.

Various studies have investigated the semi-supervised joint training of the TTS model and ASR model [1–6]. When using unpaired text data for training, a cycle called the TTS-ASR pipeline is utilized, where the multi-speaker TTS model synthesizes speech from the text with a randomly chosen reference speech and the ASR model transcribes the synthesized speech to reconstruct the source text. The conventional studies use the cycle-consistency loss [14, 15] that evaluates the distance be-

tween the source text and the reconstructed text to update the TTS model and ASR model in the pipeline. This training improves both models; the TTS model learns to correct the incorrect pronunciations so that the ASR model can accurately reconstruct the source text while the ASR model learns the speech of the unknown vocabulary and the new text sequences not included in the paired data.

However, the conventional training that simultaneously trains both models with just the cycle-consistency loss causes the TTS model to over-adapt to the ASR model, which leads to two problems. First, the synthesized speech after training does not reflect the speaker characteristics of the reference speech. We assume this is because the synthesized speech is trained to have constant speaker characteristics that the ASR model most easily recognizes. Second, pronunciation and prosody of the synthesized speech changes so that the ASR model easily recognizes them, which results in unnatural sounds for a human. These two problems not only decreases the TTS model quality but also limits the ASR model improvement, as most of the synthesized speech becomes easy examples.

To solve these problems, we propose improving the cycle-consistency-based training with a speaker consistency loss and step-wise optimization. The speaker consistency loss brings the speaker characteristics of the synthesized speech closer to that of the reference speech. Specifically, it calculates the cosine similarity between the speaker embeddings of the synthesized speech and the reference speech, which has been used to train the multi-speaker TTS model in a supervised manner [1, 16]. We propose to use this loss during the semi-supervised joint training to preserve the speaker characteristics of the synthesized speech. In the step-wise optimization, we freeze the parameter of the TTS model for a certain period at the beginning of the semi-supervised joint training. This prevents the TTS model from over-adapting to the ASR model. Our experiment shows that our proposed method achieves higher performance of both the ASR model and the TTS model compared to the conventional method.

2. Preliminaries

We utilize a Transformer-based ASR model [17] and a FastSpeech2-based TTS model [18] in this work. In this section, we describe the two models and explain how they are used in the conventional semi-supervised joint training.

2.1. Transformer-based ASR model

We denote the acoustic feature and its text as $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$ and $\mathbf{y} = (y_1, \dots, y_L)$, respectively, where $\mathbf{x}_t \in \mathbb{R}^F$ denotes the t th frame of the feature, F denotes its dimension, T denotes the frame length, y_l denotes the l th output token, and L denotes the output sequence length. In this paper, we use the time-frequency representation of the speech

as the acoustic feature and the text at phoneme level. The Transformer-based ASR model is trained to maximize the following posterior:

$$P(\mathbf{y}|\mathbf{X}; \Theta_{\text{asr}}) = \prod_l P(y_l | \mathbf{y}_{1:l-1}, \mathbf{X}; \Theta_{\text{asr}}), \quad (1)$$

where $\mathbf{y}_{1:l-1} = (y_1, \dots, y_{l-1})$ and Θ_{asr} denotes the ASR model parameter. The posterior is obtained by the encoder-decoder mechanism as follows:

$$\mathbf{H} = \text{TransformerEnc}(\mathbf{X}; \theta_{\text{asr}}^{\text{enc}}), \quad (2)$$

$$\mathbf{c}_l = \text{TransformerDec}(\mathbf{H}, \mathbf{y}_{1:l-1}; \theta_{\text{asr}}^{\text{dec}}), \quad (3)$$

$$P(y_l | \mathbf{y}_{1:l-1}, \mathbf{X}; \Theta_{\text{asr}}) = \text{softmax}(\mathbf{c}_l; \theta_{\text{asr}}^{\text{linear}})[y_l], \quad (4)$$

where $\text{TransformerEnc}(\cdot)$ is a transformer encoder that consists of a positional encoding layer and multiple multi-head self-attention blocks, $\theta_{\text{asr}}^{\text{enc}}$ denotes its parameters, $\text{TransformerDec}(\cdot)$ is a transformer decoder that consists of an embedding layer, a positional encoding layer, and multiple multi-head self-attention and encoder-decoder attention blocks, $\theta_{\text{asr}}^{\text{dec}}$ denotes its parameters, $\text{softmax}(\cdot)$ is a softmax layer with linear transformation, $\theta_{\text{asr}}^{\text{linear}}$ denotes its parameters, and $[y_l]$ denotes the element of the vector that corresponds to the probability of y_l . The parameter $\Theta_{\text{asr}} = \{\theta_{\text{asr}}^{\text{enc}}, \theta_{\text{asr}}^{\text{dec}}, \theta_{\text{asr}}^{\text{linear}}\}$ is optimized with cross-entropy loss function L_{CE} that is defined as

$$L_{\text{CE}} = -\frac{1}{L} \sum_l \log P(y_l | \mathbf{y}_{1:l-1}, \mathbf{X}; \Theta_{\text{asr}}). \quad (5)$$

2.2. FastSpeech2-based TTS model

We design our multi-speaker TTS model based on FastSpeech2 [18] to improve training speed of TTS-ASR pipeline compared to autoregressive models such as Tacotron2 [19] and TransformerTTS [20]. A reference speech is used to identify the speaker characteristics of the synthesized speech. We denote the reference speaker's speech as $\tilde{\mathbf{X}} \in \mathbb{R}^{T' \times F}$, where T' denotes the frame length of the reference speech. The estimate of the acoustic features $\hat{\mathbf{X}}$ is obtained as follows:

$$\tilde{\mathbf{s}} = \text{SpeakerModel}(\tilde{\mathbf{X}}; \theta_{\text{speaker}}), \quad (6)$$

$$\mathbf{U} = \text{FastSpeech2Enc}(\mathbf{y}; \theta_{\text{tts}}^{\text{enc}}), \quad (7)$$

$$\mathbf{V}, \hat{\mathbf{p}}, \hat{\mathbf{e}}, \hat{\mathbf{d}} = \text{VarianceAdaptor}(\mathbf{U}, \tilde{\mathbf{s}}; \theta_{\text{tts}}^{\text{va}}), \quad (8)$$

$$\hat{\mathbf{X}}_{\text{D}} = \text{FastSpeech2Dec}(\mathbf{V}; \theta_{\text{tts}}^{\text{dec}}), \quad (9)$$

$$\hat{\mathbf{X}} = \text{PostNet}(\hat{\mathbf{X}}_{\text{D}}; \theta_{\text{tts}}^{\text{post}}), \quad (10)$$

where $\text{SpeakerModel}(\cdot)$ denotes the pretrained speaker model, θ_{speaker} denotes its parameters, $\tilde{\mathbf{s}} \in \mathbb{R}^{D_s}$ denotes the speaker embedding, D_s denotes its dimension, $\text{FastSpeech2Enc}(\cdot)$ denotes a FastSpeech2 encoder that consists of an embedding layer, a positional encoding layer, and multiple multi-head self-attention blocks, $\theta_{\text{tts}}^{\text{enc}}$ denotes its parameters, $\text{VarianceAdaptor}(\cdot)$ denotes a variance adaptor that consists of a pitch predictor, a energy predictor, a duration predictor, and a length regulator, $\theta_{\text{tts}}^{\text{va}}$ denotes its parameters, \mathbf{V} denotes the phoneme hidden sequence with pitch, duration, and energy variation, $\hat{\mathbf{p}} \in \mathbb{R}^L$, $\hat{\mathbf{e}} \in \mathbb{R}^L$, and $\hat{\mathbf{d}} \in \mathbb{R}^L$ are phoneme-wise pitch, energy, and duration, respectively, $\text{FastSpeech2Dec}(\cdot)$ denotes a FastSpeech2 decoder that consists of a positional encoding layer and multiple multi-head self-attention blocks, $\theta_{\text{tts}}^{\text{dec}}$

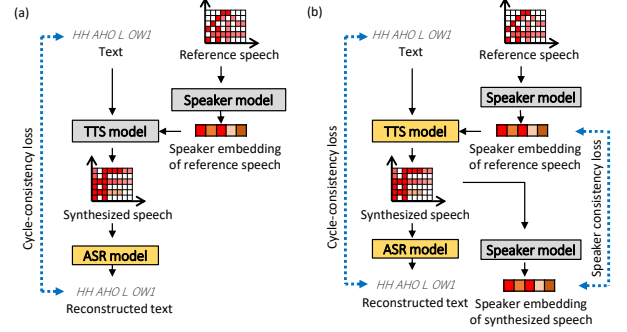


Figure 1: Overview of proposed method in (a) first-step training and (b) second-step training. Orange and gray blocks represent updated and frozen models, respectively.

denotes its parameters, $\text{PostNet}(\cdot)$ denotes a FastSpeech2 post-net, and $\theta_{\text{tts}}^{\text{post}}$ denotes its parameters. The TTS model parameter $\Theta_{\text{tts}} = \{\theta_{\text{tts}}^{\text{enc}}, \theta_{\text{tts}}^{\text{va}}, \theta_{\text{tts}}^{\text{dec}}, \theta_{\text{tts}}^{\text{post}}\}$ is optimized with the TTS loss that is defined as follows:

$$L_{\text{TTS}} = \|\mathbf{X} - \hat{\mathbf{X}}\|_1 + \|\mathbf{X} - \hat{\mathbf{X}}_{\text{D}}\|_1 + \|\mathbf{p} - \hat{\mathbf{p}}\|_2^2 + \|\mathbf{e} - \hat{\mathbf{e}}\|_2^2 + \|\mathbf{d} - \hat{\mathbf{d}}\|_2^2, \quad (11)$$

where $\|\cdot\|_1$ denotes the L1 norm, $\|\cdot\|_2$ denotes the L2 norm, and \mathbf{p} , \mathbf{e} , and \mathbf{d} denote the ground-truth pitch, energy, and duration, respectively.

2.3. Conventional semi-supervised joint training

In this section, we describe the baseline framework for semi-supervised joint training of TTS and ASR using paired data and unpaired text data. First, the TTS model and the ASR model are separately pretrained with paired data using (5) and (11). Then, for cycle-consistency training utilizing unpaired text data and paired data, a TTS-ASR pipeline is created, in which the multi-speaker TTS model synthesizes speech from unpaired text data using a random reference speech sampled from the paired data and the ASR model transcribes the synthesized speech to reconstruct the source text [1–6]. The TTS-ASR pipeline is optimized with the cycle-consistency loss that is defined as

$$L_{\text{cycle}} = -\frac{1}{L} \sum_l \log P(y_l | \mathbf{y}_{1:l-1}, \hat{\mathbf{X}}; \Theta_{\text{asr}}), \quad (12)$$

$$\hat{\mathbf{X}} = \text{TTS}(\mathbf{y}, \tilde{\mathbf{X}}; \Theta_{\text{tts}}), \quad (13)$$

where $\text{TTS}(\cdot)$ is a function that combines Eqs. (6)–(10). Note that we omit $\hat{\mathbf{p}}$, $\hat{\mathbf{e}}$, and $\hat{\mathbf{d}}$ because they are not used in the semi-supervised joint training.

3. Proposed method

3.1. Strategy

In conventional semi-supervised training utilizing unpaired text data, the TTS model and the ASR model are jointly trained with cycle-consistency loss (12). However, as discussed in section 1, the speaker characteristics of the synthesized speech is not preserved in this case and the synthesized speech becomes overly easy for the ASR model to recognize, which leads not only to a low-quality TTS model but also to a limitation of the ASR improvement. Our proposed method addresses this problem by utilizing step-wise optimization and the speaker consistency loss.

The overview of the proposed method is shown in Fig. 1. Our proposed step-wise optimization trains the ASR model and the TTS model in two steps. In the first step (Fig. 1(a)), we fix the parameter of the TTS model and optimize the ASR model. In the second step (Fig. 1(b)), we train both the TTS model and the ASR model. The key point of this step-wise optimization is that the ASR model is optimized with the synthesized speech before the TTS model and ASR model are trained. When the ASR model is not trained with the synthesized speech, the cycle-consistency loss becomes large at the beginning of the semi-supervised joint training, which drives the TTS model to synthesize speech that is easy for the ASR model to recognize. In contrast, since the ASR model is trained with the synthesized speech after the first step of the proposed step-wise optimization, over-adaptation of the TTS model to the ASR model is prevented. We experimentally show it in section 4.4.

The speaker consistency loss calculates the cosine similarity between the embeddings of the synthesized speech and that of the reference speech (shown in Fig. 1 (b)), and brings the speaker characteristics of the synthesized speech close to those of the reference speech. This loss is beneficial for keeping the speaker characteristics of the synthesized speech.

3.2. Training

In this section, we formulate the speaker consistency loss and describe the step-wise optimization. Given the synthesized speech \hat{X} , we estimate its speaker embedding as

$$\hat{s} = \text{SpeakerModel}(\hat{X}; \theta_{\text{speaker}}), \quad (14)$$

where the speaker model and its parameters are the same as those in (6). We define the speaker consistency loss L_{sc} as

$$L_{sc} = -\frac{\hat{s}^T \tilde{s}}{\|\hat{s}\|_2 \|\tilde{s}\|_2}, \quad (15)$$

where T denotes the transpose of a vector. Minimizing the speaker consistency loss maximizes the similarity between embedding of the synthesized speech and that of the reference speech, which preserves the speaker characteristics of the synthesized speech.

We formulate the step-wise optimization with the speaker consistency loss as follows. First, we freeze the parameter of the TTS model in the TTS-ASR pipeline and train the ASR model with the cycle-consistency loss (12). Then, we unfreeze the parameter of the TTS model and train both the ASR model and the TTS model with the cycle-consistency loss and the speaker consistency loss, which is defined as

$$L_{\text{prop}} = L_{\text{cycle}} + \alpha L_{sc}, \quad (16)$$

where α is the loss weight.

4. Experiment

4.1. Dataset

We used the VoxCeleb2 dataset [21] for the speaker model training and the LibriTTS dataset [22] for the TTS model and ASR model training and evaluation. The speaker model was pre-trained with a speaker classification of 5,994 speakers using the dev set of VoxCeleb2. The `train-clean-100` set (train-100) and the `train-clean-360` set (train-360) of LibriTTS was used as paired data and unpaired text data, respectively.

We used `dev-clean` set and `test-clean-100` set of LibriTTS as validation data and test data, respectively. We used 80 log mel-scale filterbank coefficients as acoustic features, which were extracted using a 50-ms-long Hann window with a 12.5-ms-long shift. All the text was converted to phonemes and we inserted a word space token between each word. To train FastSpeech2, we estimated the ground-truth duration, pitch, and energy of paired data with the Montreal Forced Aligner [23], PyWorld [24, 25], and L2 norm of each frame following [26]. We used speech data with less than 900 frames and text data with less than 180 tokens for the memory constraint. The sampling rate of all data was 16 kHz.

4.2. Implementation

The speaker model consisted of three bidirectional LSTM (BLSTM) layers with 256 units. The forward and backward outputs of the BLSTM were concatenated. The third BLSTM output was averaged with an attention mechanism to obtain the target speaker embedding, which calculated the weighted mean of frame-level features [27]. The TTS model consisted of four encoder layers and six decoder layers. We performed the phoneme-level prediction of energy and pitch as in [26] and added the speaker embedding from the speaker model to the encoder outputs. The post-net was the same as that of Tacotron2 [19]. The other settings were the same as that in [18]. The ASR model consisted of six encoder layers and four decoder layers. The dimension of each transformer block was set to 512 and the number of attention heads was four.

4.3. Settings

We compared the following methods (listed in Table 1): pretrained model trained with paired train-100, conventional method trained with paired train-100 and unpaired train-360, the proposed method trained with paired train-100 and unpaired train-360, and full-supervised model trained with paired train-100 and paired train-360. To determine the effect of the speaker consistency loss and step-wise optimization, we also conducted an ablation study on each method. The speaker model was optimized using the Adam [28] algorithm with a minibatch size of 128. We set the learning rate of the algorithm to $1e - 3$. The training steps were stopped if the loss on the validation set did not decrease for five epochs in succession. We fixed the parameter of the speaker model after the pretraining. The TTS model was optimized in the same way as the speaker model except that we set the minibatch size to 32 and the learning rate to $1e - 4$. The ASR model was optimized using the RAdam [29] algorithm with a minibatch size of 32. The learning rate of the algorithm was set to $1e - 4$. Early stopping was used in the same way as the speaker model. We used scheduled sampling [30]-based optimization during the ASR training, where teacher forcing is used at the beginning of training, and we linearly increased the probability of sampling to the probability of 0.4 at 20 epochs. We also applied the time-masking and frequency-masking of SpecAugment [31], where the number of time-masks and frequency-masks is both two and the masking width is randomly set between 0 and 100 for the time-masking and between 0 and 27 for the frequency-masking.

For the semi-supervised joint training, we used the RAdam algorithm with a minibatch size of 8. The learning rate was set to $1e - 5$. We set α in (16) to 0.1. We trained the ASR model until convergence in the first step of the step-wise optimization. We mixed the supervised paired data from train-100 and the unsupervised text data from train-360 and provided the

Table 1: Evaluation results

Method	Speaker consistency	Step-wise optimization	Dataset	PER (%)	MCD (dB)	F0 RMSE (Hz)
Pretrained model			Paired train-100	15.5	8.3	36.7
Conventional method			Paired train-100 & unpaired train-360	13.0	7.2	43.1
Proposed method	✓	✓	Paired train-100 & unpaired train-360	9.6	6.9	33.2
w/o speaker consistency		✓	Paired train-100 & unpaired train-360	9.6	7.5	39.8
w/o step-wise optimization	✓		Paired train-100 & unpaired train-360	12.5	6.9	34.8
Full-supervised model			Paired train-100 & paired train-360	6.6	7.7	37.0

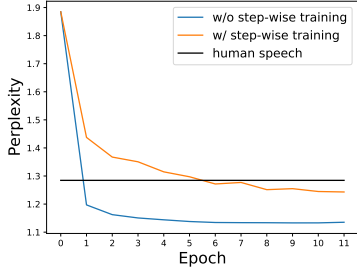


Figure 2: Perplexity curve calculated using synthesized speech and human speech. Blue and orange lines represent perplexity calculated using synthesized speech trained without and with step-wise optimization, respectively. Black line represents perplexity calculated using human speech.

mixed data randomly to the models. When unsupervised text data was given, the models were optimized using (12) or (16). Otherwise, the models were optimized using (5) and (11). We applied the time-masking and frequency-masking of SpecAugment to the synthesized speech in the same way as those during ASR pretraining. The reference speech for the TTS model was randomly chosen from the paired data. Since the parameter of the duration predictor of the TTS model was not differentiable, we fixed this parameter during training. We used phoneme error rate (PER) as an ASR evaluation metric, and mel-cestral distortion (MCD) [32] and root mean square error (RMSE) of fundamental frequency (F0) as the TTS evaluation metrics. Dynamic time warping was used to calculate MCD and F0 RMSE. We used MelGAN [33] as a vocoder and calculated F0 with DIO and StoneMask of PyWorld [24, 25]. To clarify the effect of the proposed step-wise optimization, we plotted the perplexity of the synthesized speech trained with and without step-wise optimization using validation data, which is calculated as

$$\text{perp.} = \exp \left(-\frac{1}{L} \sum_l \ln P(y_l | \mathbf{y}_{1:l-1}, \hat{\mathbf{X}}; \Theta_{\text{ASR}}^{\text{pre}}) \right), \quad (17)$$

where $\Theta_{\text{ASR}}^{\text{pre}}$ denotes the parameter of the pretrained ASR model that is not trained with the synthesized speech. The low perplexity indicates that the ASR model accurately estimates the posterior from the synthesized speech.

4.4. Results

Table 1 shows the result of each method. As we can see, the proposed method outperforms the conventional method in terms of both ASR and TTS performance. Specifically, the F0 RMSE of the conventional method is worse compared to that of the pre-trained model. This is because the semi-supervised joint training with just the cycle-consistency loss drives the TTS model to synthesize speech of the speaker that the ASR most easily recognizes. In contrast, both MCD and F0 RMSE of the proposed method improve over the conventional method thanks to

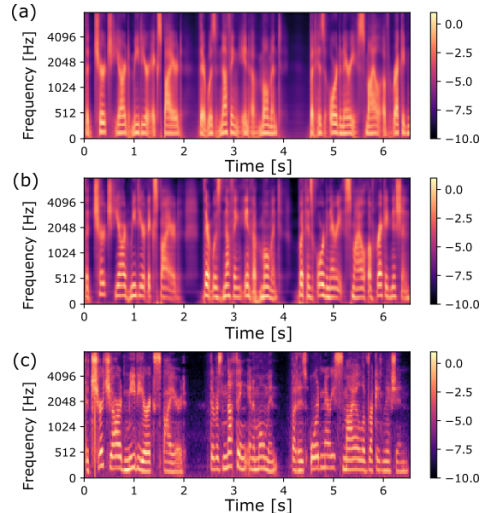


Figure 3: Mel-spectrograms of (a) conventional method, (b) proposed method, and (c) ground truth.

introducing the speaker consistency loss. Table 1 also shows that the PER of the ASR model improves with our proposed step-wise optimization. Figure 2 shows the perplexity curve of the proposed method with and without step-wise optimization. We also plotted the perplexity calculated with the human speech as reference. We can see here that the perplexity of the TTS model trained without step-wise optimization becomes far lower compared to human speech after just one epoch, which indicates that the TTS model over-adapts to the ASR model and the synthesized speech becomes overly easy for the ASR model to recognize. In contrast, the perplexity curve of the TTS model trained with the proposed step-wise optimization becomes gentle, and perplexity converges to the point closer to that calculated with human speech, which indicates the proposed step-wise optimization prevents the over-adaptation of the TTS model to the ASR model. Figure 3 shows the mel-spectrograms of the conventional method, the proposed method, and the ground truth. The proposed method reflected the characteristics of the ground-truth speech that the conventional method failed to reflect.

5. Conclusion

In this paper, we proposed to improve the cycle-consistency-based training with the speaker consistency loss and step-wise optimization. The speaker consistency loss brings the speaker characteristics of the synthesized speech closer to that of the reference speech. The step-wise optimization prevents the over-adaptation of the TTS model. Experimental results showed that the speaker consistency loss improves TTS quality in terms of MCD and F0 RMSE, and the step-wise optimization prevents over-adaptation of the TTS model, which leads to the improvement of both the ASR model and the TTS model.

6. References

- [1] A. Tjandra, S. Sakti, and S. Nakamura, “Machine speech chain,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 976–989, 2020.
- [2] S. Karita, S. Watanabe, T. Iwata, M. Delcroix, A. Ogawa, and T. Nakatani, “Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders,” in *Proc. ICASSP*, 2019, pp. 6166–6170.
- [3] T. Hori, R. F. Astudillo, T. Hayashi, Y. Zhang, S. Watanabe, and J. L. Roux, “Cycle-consistency training for end-to-end speech recognition,” in *Proc. ICASSP*, 2019, pp. 6271–6275.
- [4] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, “Almost unsupervised text to speech and automatic speech recognition,” in *Proc. ICML*, vol. 97, 2019, pp. 5410–5419.
- [5] J. Xu, X. Tan, Y. Ren, T. Qin, J. Li, S. Zhao, and T. Liu, “Lr-speech: Extremely low-resource speech synthesis and recognition,” in *Proc. KDD*, 2020, pp. 2802–2812.
- [6] M. K. Baskar, L. Burget, S. Watanabe, R. F. Astudillo, and J. H. Cernocky, “Eat: Enhanced ASR-TTS for self-supervised speech recognition,” in *Proc. ICASSP*, 2021, pp. 6753–6757.
- [7] Y. Chung, Y. Wang, W. Hsu, Y. Zhang, and R. J. Skerry-Ryan, “Semi-supervised training for improving data efficiency in end-to-end speech synthesis,” in *Proc. ICASSP*, 2019, pp. 6940–6944.
- [8] M. Hwang, R. Yamamoto, E. Song, and J. Kim, “TTS-by-TTS: TTS-driven data augmentation for fast and high-quality speech synthesis,” in *Proc. ICASSP*, 2021, pp. 6598–6602.
- [9] Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu, “Png BERT: Augmented BERT on phonemes and graphemes for neural TTS,” in *Proc. INTERSPEECH*, 2021, pp. 151–155.
- [10] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, “An analysis of incorporating an external language model into a sequence-to-sequence model,” in *Proc. ICASSP*, 2018, pp. 5824–5828.
- [11] A. Sriram, H. Jun, S. Satheesh, and A. Coates, “Cold fusion: Training seq2seq models together with language models,” in *Proc. INTERSPEECH*, 2018, pp. 387–391.
- [12] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. J. Moreno, Y. Wu, and Z. Wu, “Speech recognition with augmented synthesized speech,” in *Proc. ASRU*, 2019, pp. 996–1002.
- [13] P. B. Denes and E. N. Pinson, *The Speech Chain*. Macmillan, 1993.
- [14] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T. Liu, and W. Ma, “Dual learning for machine translation,” in *Proc. NIPS*, 2016, pp. 820–828.
- [15] J. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. ICCV*, 2017, pp. 2242–2251.
- [16] Z. Cai, C. Zhang, and M. Li, “From speaker verification to multi-speaker speech synthesis, deep transfer with feedback constraint,” in *Proc. INTERSPEECH*, 2020, pp. 3974–3978.
- [17] L. Dong, S. Xu, and B. Xu, “Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition,” in *Proc. ICASSP*, 2018, pp. 5884–5888.
- [18] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, “FastSpeech 2: Fast and high-quality end-to-end text to speech,” in *Proc. ICLR*, 2021.
- [19] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyriannakis, and Y. Wu, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, 2018, pp. 4779–4783.
- [20] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, “Neural speech synthesis with transformer network,” in *Proc. AAAI*, 2019, pp. 6706–6713.
- [21] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep speaker recognition,” in *Proc. INTERSPEECH*, 2018, pp. 1086–1090.
- [22] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: A corpus derived from librispeech for text-to-speech,” in *Proc. INTERSPEECH*, 2019, pp. 1526–1530.
- [23] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using kaldii,” in *Proc. INTERSPEECH*, 2017, pp. 498–502.
- [24] M. Morise, “CheapTrick, a spectral envelope estimator for high-quality speech synthesis,” *Speech Communication*, vol. 67, pp. 1–7, 2015.
- [25] —, “Error evaluation of an F0-adaptive spectral envelope estimator in robustness against the additive noise and F0 error,” *IEICE Trans. on Information and Systems*, vol. 98-D, no. 7, pp. 1405–1408, 2015.
- [26] C. Chien, J. Lin, C. Huang, P. Hsu, and H. Lee, “Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech,” in *Proc. ICASSP*, 2021, pp. 8588–8592.
- [27] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive statistics pooling for deep speaker embedding,” in *Proc. INTERSPEECH*, 2018, pp. 2252–2256.
- [28] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. ICLR*, 2015.
- [29] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, “On the variance of the adaptive learning rate and beyond,” in *Proc. ICLR*, 2020.
- [30] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, “Scheduled sampling for sequence prediction with recurrent neural networks,” in *Proc. NIPS*, 2015, pp. 1171–1179.
- [31] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” in *Proc. INTERSPEECH*, 2019, pp. 2613–2617.
- [32] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *Proc. PACRIM*, vol. 1, 1993, pp. 125–128.
- [33] K. Kumar, R. Kumar, T. de Boissiere, L. Gestein, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, “Mel-GAN: Generative adversarial networks for conditional waveform synthesis,” in *Proc. NeurIPS*, 2019.