INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Dynamic Speaker Representations Adjustment and Decoder Factorization for Speaker Adaptation in End-to-End Speech Synthesis

Ruibo Fu[1,2], *Jianhua Tao*[1,2,3], *Zhengqi Wen*[1], *Jiangyan Yi*[1], *Tao Wang*[1,2], *Chunyu Qiang*[1,2]

[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing
[3]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing

{ruibo.fu,jhtao,zqwen,jiangyan.yi,tao.wang,chunyu.qiang}@nlpr.ia.ac.cn

Abstract

End-to-end speech synthesis can reach high quality and naturalness with low-resource adaptation data. However, the generalization of out-domain texts and the improving modeling accuracy of speaker representations are still challenging tasks. The limited adaptation data leads to unacceptable errors and low similarity of the synthetic speech. In this paper, both speaker representations modeling and acoustic model structure are improved for the speaker adaptation task. On the one hand, compared with the conventional methods that focused on using fixed global speaker representations, the attention gating is proposed to adjust speaker representations dynamically based on the attended context and prosody information, which can describe more pronunciation characteristics in phoneme level. On the other hand, to improve the robustness and avoid over-fitting, the decoder model is factored into average-net and adaptation-net, which are designed for learning speaker independent acoustic features and target speaker timbre imitation respectively. And the context discriminator is pre-trained by large ASR data to supervise the average-net generating proper speaker independent acoustic features for different phoneme. Experimental results on Mandarin dataset show that proposed methods lead to an improvement on intelligibility, naturalness and similarity.

Index Terms: speech synthesis, speaker adaptation, dynamic speaker representations adjustment, decoder factorization

1. Introduction

End-to-end speech synthesis, such as Tacotron, can achieve the state-of-art performance, and even close to human recording based on a large corpus [1–5]. However, in the circumstance of limited target speaker adaptation data, the generalization of out-domain texts is still a challenge. A lot of unacceptable errors could occur, including skipping, repeating, mispronunciation, and etc. Besides, the speaker representations used to model timbre discrepancy is difficult for extraction, which is inaccurate and lack of timbre control ability on the synthetic speech.

Generally, speaker adaptive training methods boil down to two aspects. One aspect is the speaker representations. A type of speaker representations modeling methods is based on the speaker recognition task, where i-vectors [6–8], d-vectors [9] and speaker encoder networks [10] are used in the acoustic model to control speech generation styles. This kind of method can build a new speaker style fast without training procedure. But above speaker representations used in the speaker recognition task is not optimal for the speech synthesis task. Besides, both two tasks need a large and variable corpus to model the sparse speaker representations space. Another type of speaker representations modeling methods is to use trainable embedding as one of the condition inputs. Researchs such as Global Style Token [11] and VoiceLoop [12, 13] use additional reference audio as input to acquire speaker styles. And some other researches, such Deep Voice [2, 14], use one-hot to distinguish different speakers and generate speaker embedding through nonlinear transformer [15]. All the above methods assume that a fixed dimension context independent speaker representation could be obtained by training. However, context differences would cause the disturbance of speaker representations, which effects the modeling accuracy.

Another aspect for speaker adaptive training is the model structure. In the traditional pipeline speech synthesis framework [16], to achieve adaptation in various network for each style of speaker, most of researches applied speaker dependent layers [17, 18] and speaker and language factorization methods [19–21]. In the end-to-end speech synthesis framework, most of speaker adaptation researches [14, 15] are based on Tacotron [1, 22], which is an attention based encoder-decoder structure. Unlike pipeline framework, the decoder model only relied on speaker representations as condition input to guide various style generations. But the whole decoder model may occur over-fitting phenomenon when the adaptation database is limited, which could lead to unacceptable errors in the speech synthesis process.

In this paper, both speaker representations modeling and acoustic model structure are improved for the speaker adaptation task. First, to model the speaker representations distortion caused by different context, the attention gating is proposed to adjust speaker representations dynamically based on the attended context information and prosody information, which is extracted from attention alignments results. Second, to improve the robustness of synthetic speech and avoid over-fitting phenomenon, the decoder model structure is factored into average-net and adaptation-net. The decoder average-net is designed for learning speaker independent intermediate acoustic features, which contain the basic context pronunciation information in the acoustic level. And the context discriminator is pre-trained by large ASR data to supervise the average-net generating proper speaker independent acoustic features in different attended context. Meanwhile, the decoder adaptation-net are designed for learning target speaker timbre imitation condition by the speaker representations.

Overall, the contributions of this paper are two-fold. First, the attention gating is proposed to shift the speaker representations dynamically based on the context and prosody. Second, decoder model factorization is proposed to learn pronunciation and timbre separately. Experiments demonstrate the robustness and similarity improvements by applying proposed methods.

The rest of the paper is organized as follows. Section 2 describes methods. Experiments and results are analyzed in section 3 and 4. The conclusions are discussed in Section 5.

http://dx.doi.org/10.21437/Interspeech.2020-1623

2. Method

Fig.1 shows the architecture of the Tacotron based speaker adaptation end-to-end framework. The whole network consists of two components.

In the acoustic model part, the whole seq-to-seq model consists of three parts: Encoder mainly processes text information. The attention mechanism connects the encoder and decoder and controls the prosody. Based on our previous work [23], we add an extra prosody embedding in the attention mechanism to model the duration distribution, which is conditioned on the context information. And the attended phoneme and duration information for each decoder time step could be obtain from the attention alignment results. For each decoder time step i, the attention gating generates local shift embedding $E_{LS}^{(i)}$ base on the phoneme, duration, prosody and global speaker embeddings. After the shifting procedure, the final context and prosody dependent shifted speaker representation $E_S^{(i)}$ would be fed into the decoder adaptation-net to model acoustic variation between different speakers. The decoder generates acoustic features conditioned on speaker embeddings to ensure high similarity of the target speaker. The decoder average-net is designed for learning speaker independent intermediate acoustic features, which is the input of context discriminator. The context discriminator is pre-trained with the whole acoustic seq-to-seq model by large ASR data, which aims to supervise the average-net generating speaker independent acoustic features.

In the neural vocoder part, we deploy the LPCNet [24], which significantly improves the efficiency of speech synthesis and remains high quality. In the following sections, the dynamic speaker representations adjustment and decoder model factorization would be introduced.

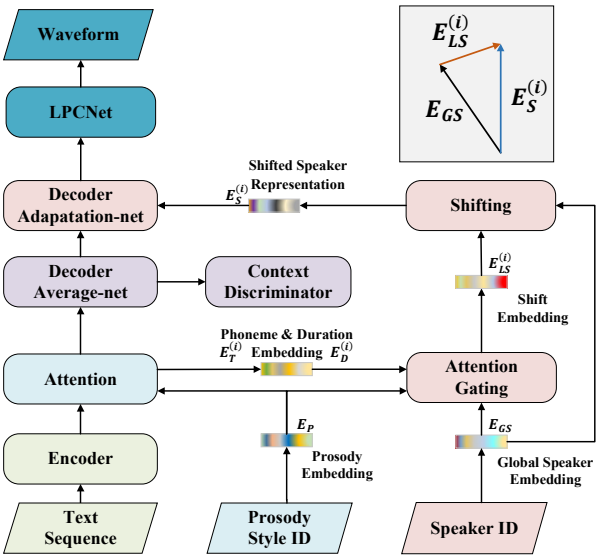

Figure 1: *System architecture of the Tacotron based speaker adaptation end-to-end framework. Before the shifting procedure generates the shifted speaker representation $E_S^{(i)}$, the attention gating generates shift embedding $E_{LS}^{(i)}$ based on the context and prosody information. The decoder model is factored into average-net and adaptation-net, which are designed for learning speaker independent acoustic features and target speaker timbre imitation respectively.*

2.1. Dynamic Speaker Representations Adjustment

The dynamic speaker representations adjustment mechanism is designed to generate context and prosody dependent speaker representations for each decoder time step. As shown in the Fig.1, the attention gating network computes the local shift embedding by learning a non-linear combination among context embedding, duration embedding and prosody embedding. Our key insight is that for each decoder time step depending on the attended context, the speaker representations used to guide the decoder may differ. For instance, the distribution of acoustic features in the circumstance of vowel and consonant is uninformative. To handle these dynamic dependencies better, the gating mechanism is proposed to control the importance of each embedding.

The inputs of attention gating is defined as followed: The whole context input sequence is $X = (X_1, X_2, ..., X_U)$ with U dimension. For each decoder time i, similar with our previous work [23], the attention weights $\alpha_{i,j}$, which is also called as alignments, is calculated by the following equation:

$$\alpha_{i,j} = \text{Attention}\left(s_i, h_j, \alpha_{i-1}, E_P\right) \tag{1}$$

where s_i denotes state, h_j denotes query,$j = 1, \ldots, U$, and E_P denotes the prosody embedding for modeling different prosody style. The index of current attended phoneme in the input sequence x is $T_{index}^{(i)}$:

$$T_{index}^{(i)} = \arg \max_{1 \le j \le U} \left(\alpha_{i,j}\right) \tag{2}$$

Therefore, the current attended phoneme embedding $E_T^{(i)}$ could be built by nonlinear transform according to the current phoneme $X(T_{index}^{(i)})$. The duration $D^{(i)}$ is defined as the continuing decoder steps in the current attended phoneme, which is defined by the following equation:

$$D^{(i)} = \begin{cases} D^{(i-1)} + 1, & if \quad T_{index}^{(i)} = T_{index}^{(i-1)} \\ 0, & if \quad T_{index}^{(i)} \neq T_{index}^{(i-1)} \end{cases} \tag{3}$$

Then the duration embedding $E_D^{(i)}$ could be built by nonlinear transform according to $D^{(i)}$. Besides, the global speaker embedding E_{GS} could also be built according to input speaker ID, which is the fixed part of speaker representations for each speaker. All the above embeddings are initialized with Glorot [25] initialization.

The context gate $G_C^{(i)}$ and prosody gate $G_P^{(i)}$ are defined by the following equations:

$$G_C^{(i)} = \sigma\left(w_C \left[E_T^{(i)}; E_D^{(i)}; E_{GS}\right] + b_C\right) \tag{4}$$

$$G_P^{(i)} = \sigma\left(w_P \left[E_P; E_{GS}\right] + b_P\right) \tag{5}$$

where $[;]$ denotes the operation of vector concatenation, w_C and w_P are weight vectors, b_C and b_P are biases, $\sigma(\cdot)$ represents sigmoid function.

Then the local context and prosody dependent shift embedding $E_{LS}^{(i)}$ is calculated by the following equation:

$$E_{LS}^{(i)} = G_C^{(i)}\left(w_{hc}\left[E_T^{(i)}; E_D^{(i)}\right]\right) + G_P^{(i)}\left(w_{hp} E_P^{(i)}\right) + b_h^{(i)} \tag{6}$$

where w_{hc} and w_{hp} are weight vectors, $b_h^{(i)}$ is bias vector.

After the shifting procedure, the final shifted context and prosody dependent speaker representation $E_S^{(i)}$ is calculated by the following equation:

$$E_S^{(i)} = E_{GS} + \alpha E_{LS}^{(i)} \tag{7}$$

$$\alpha = \min\left(\frac{\|\mathbf{E}_{GS}\|_2}{\left\|\mathbf{E}_{LS}^{(i)}\right\|_2}\beta, 1\right) \tag{8}$$

where β is the hyper-parameter. In order to avoid the magnitude of the shifted embedding $E_{LS}^{(i)}$ is too large compared with the global speaker embedding E_{GS}, the scaling factor α is designed to constrain the magnitude of the shift embedding.

2.2. Decoder Model Factorization

Last section introduces the dynamic speaker representations adjustment mechanism, the final context and prosody dependent shifted speaker representation $E_S^{(i)}$ would be fed into the decoder model. Compared with traditional methods that feed the speaker representation into the whole encoder model to conduct speaker adaptation training, the decoder model is factored into adaptation-net and average-net. Our key insight is that a more speaker independent intermediate acoustic features could be learned by a large multi-speaker database in the decoder average-net, which could ensure the intelligibility and robustness of the synthetic speech. And only a part of decoder, which is the average-net, is built for modeling multi-style speech synthesis to avoid over-fitting phenomenon and performance degradation in the out-of-domain text.

To ensure the average-net generating speaker independent acoustic features, the context discriminator is pre-trained with the whole acoustic seq-to-seq model by large ASR data. To be more specific, the designed model is based on the Tacotron 2 [1], the decoder average-net consists of the Pre-Net, which is two fully connected layers, and one layer of LSTM. The output of average-net is the input of adaptation-net and context discriminator. The context discriminator consists of three fully connected layers with a sigmoid output layer, which predicts the phoneme classification results for each decoder time step. The cross entropy error between the predicted labels and $X(T_{index}^{(i)})$ analyzed from the attention alignment results is defined as one part of the total training loss. In the pre-training process, the context discriminator would be optimized jointly with the acoustic model. In the adaptive training process, the parameters of context discriminator would be frozen. Based on the intermediate acoustic features from average-net, the adaptation-net generates acoustic features conditioned on shifted speaker representation $E_S^{(i)}$ to ensure high similarity of the target speaker. The adaptation-net consists of one layer of LSTM, linear projection layer and stop token layer.

3. Experimental Setup

We use the Blizzard Challenge 2019 dataset and our own internal dataset to conduct the experiments. Our internal dataset consists of 25 different professional Mandarin speakers with about 200 hours. The Blizzard Challenge dataset is an estimated 8 hours of speech from one native Mandarin speaker collected from talk shows. One male and one female speaker are recorded with about 20 minutes for the speaker adaptative training. All the wav files are sampled at 16kHz. Besides, the AISHELL-1 ASR data is also used for the pre-training of the context discriminator [26]. In this work, we limit the input of the synthesis to 32 features: The 30-dim Bark-scale [27] cepstral coefficients, and 2 pitch parameters (period, correlation) are extracted directly from recorded speech samples. The input text is processed by our G2P frontend and transformed to the phoneme sequences, which also include Mandarin tone information of vowels.

For the Tacotron training, we set output layer reduction factor $r = 2$. And we use the GMM attention mechanism described in the research [28], where the prosody and duration embedding are added to predict the attention weights. We use the Adam optimizer with adaptative learning rate decay, which starts from 0.0001 and decays as introduced in our previous research [23]. The training batch size is 16, where all sequences are padded to a max length. There are about 600K global steps for the pre-training of context discriminator and the whole acoustic seq-to-seq model by the AISHELL-1 ASR data. For the low-resource speaker adaptation task, after about 600K global steps by using the multi-speaker TTS data, there are about 2-3K global steps for adaptative training.

For the LPCNet training, the network is trained for 120 epochs, with a batch size of 64, each sequence consisting of 15 10-ms frames. We use the AMSGrad [29] optimization method (Adam variant) with a step size $\alpha = \alpha_0/(1 + \delta \cdot b)$ where $\alpha_0 = 0.001$, $\delta = 5 \times 10^{-5}$, and b is the batch number. For the LPCNet adaptation, there are about 10 epochs for adaptative training.

The models on which we conduct experiments include:

- **SI-Base:** The speaker independent Tacotron2 baseline model is trained by multi-speaker data but without speaker identity information [1]. The speaker adaptation data is used to fine-tune the average model trained by the multi-speaker data.
- **SD-Base:** The speaker dependent Tacotron2 baseline model use one-hot to distinguish different speakers and generate speaker embedding through nonlinear transformer [15].
- **P-*-*:** Our proposed method is denoted as P. To do ablation studies, we make several models. SSE and FSE is short for proposed shifted speaker embedding method and baseline fixed speaker embedding method respectively. FD and MD are short for proposed factored decoder model and baseline mono decoder model respectively. For instance, the P-SSE-FD system is our final proposed method.

We evaluate the performance of our models in terms of intelligibility, naturalness and similarity. The test sets are about 500 utterances, containing the in-domain and out-domain text, which involve news, encyclopedias, story and poetry. To evaluate intelligibility, a subset for about 50 utterances is selected by sorting high frequency unacceptable errors based on baseline evaluations. 30 listeners conducted crowd-sourcing ABX preference tests and MOS tests. In each experimental group, 30 parallel sentences are selected randomly from test subset.

4. Evaluation and Discussion

4.1. Convergence Speed and Intelligibility Evaluation

First, the convergence speed is mainly evaluated based the attention alignments results. As shown in the Fig.2, two systems (P-SEE-MD/FD) are compared in the speaker adaptation task. The proposed decoder model factorization method could faster

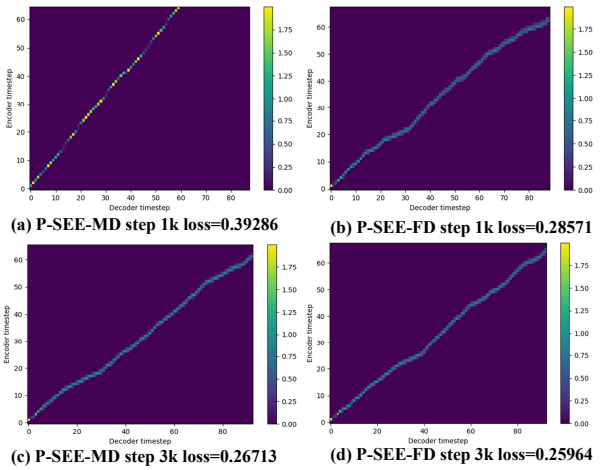

Figure 2: *Attention alignments and loss results with the same text on a test utterance in different systems.*

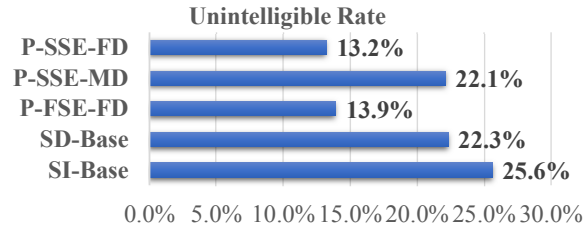

Figure 3: *Utterance level intelligibility results.*

the convergence speed and improve the accuracy of the acoustic model. Especially in the early adaptation training stage, the decoder average-net could speed up the proper matching between the contexts and audio acoustic features. We infer that the introduction of context discriminator could guide the better generation of intermediate acoustic features and faster the convergence of trained model.

Intelligibility tests are performed with metrics of case level unintelligible rate to evaluate the robustness of models. If the synthetic speech utterance contains the unacceptable errors such as skipping, repeating and mispronunciation, this unintelligible utterance would be counted. The utterance level intelligibility results are shown in the Fig.3. We could observe that the decoder model factorization method decreases the unintelligible rate about 40%, in which the decoder average-net plays an important role in improving the robustness of the system. Besides, by comparing the unintelligible rate of P-SSE/FSE-FD, the dynamic speaker representations adjustment method could further improve the robustness. By analyzing the synthetic speech, we find that this further improvement mainly involves phoneme related mispronunciations. It can be interpreted that the context dependent speaker representations could guide the decoder generate more accurate acoustic features, which could avoid some disturbance from the noise in the fixed speaker representations.

4.2. Naturalness and Similarity Evaluation

In this part, we first evaluate the naturalness of the synthetic speech from different models. The ABX test results on the naturalness is shown in the Tab.1. By observing the preference scores, the proposed P-SSE-FD system achieve better performance than the baseline systems. Besides, the decoder model factorization method makes more contributions on naturalness improvement than the dynamic speaker representations adjustment method according to the ablation results. A possible explanation is that people would more focusing on the unacceptable errors in the naturalness evaluation, which is more related to the robustness.

The similarity of the synthetic speech is one of key measures in the low-resource speaker adaptation task. The MOS results on similarity of synthetic speech is illustrated in the Fig.4. The male and female speaker adaptation model is evaluated separately. The female speaker adaptation model performs better in all the systems. It can be interpreted that the female speaker data is larger than the male speaker data in the training process of average model. Besides, we observe that the proposed dynamic speaker representations adjustment method could increase the similarity MOS for about 0.4 point.

Table 1: *Preference scores on naturalness of synthetic speech.*

System A	Scores A (%)	Scores Neutral (%)	Scores B (%)	System B
P-SSE-FD	62.72	13.45	23.83	SI-Base
	58.25	12.37	29.38	SD-Base
	53.83	18.46	27.71	P-SSE-MD
	47.92	15.41	36.67	P-FSE-FD

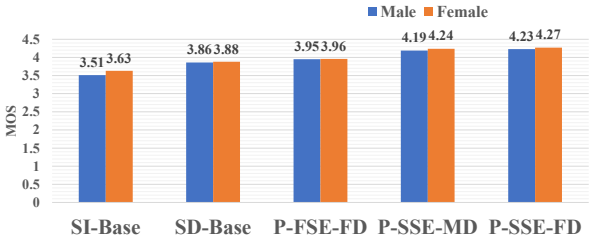

Figure 4: *MOS results on similarity of synthetic speech.*

5. Conclusions

In this paper, we propose a dynamic speaker representations adjustment mechanism and decoder factorization for speaker adaptation in end-to-end speech synthesis system. The shifted speaker embedding can consider phoneme level acoustic features discrepancy and improve the modeling accuracy. The decoder factorization can realize the functions decomposition of pronunciation construction and style learning. Experimental results demonstrate that both the methods improve intelligibility, naturalness and similarity of the synthetic speech.

6. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2017YFB1002801), the National Natural Science Foundation of China (NSFC) (No.61831022, No.61901473, No.61771472, No.61773379) and the Major Program for the National Social Science Fund of China (13&ZD189). This work is also supported the CCF-Tencent Open Research Fund.

7. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. 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.

[2] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *ICLR*, 2017.

[3] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality tts with transformer," *arXiv: Computation and Language*, 2018.

[4] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[5] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International Conference on Machine Learning*, 2018, pp. 3918–3926.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[7] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for dnn-based speech synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] Y. Zhao, D. Saito, and N. Minematsu, "Speaker representations for speaker adaptation in multiple speakers' blstm-rnn-based speech synthesis," in *Interspeech 2016*, 2016, pp. 2268–2272.

[9] R. Doddipatla, N. Braunschweiler, and R. Maia, "Speaker adaptation in dnn-based speech synthesis using d-vectors." in *INTERSPEECH*, 2017, pp. 3404–3408.

[10] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in neural information processing systems*, 2018, pp. 4480–4490.

[11] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*, 2018, pp. 5180–5189.

[12] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," *arXiv preprint arXiv:1707.06588*, 2017.

[13] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," in *International Conference on Machine Learning*, 2018, pp. 3683–3691.

[14] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in neural information processing systems*, 2017, pp. 2962–2970.

[15] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *International Conference on Machine Learning*, 2018, pp. 4693–4702.

[16] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *speech communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[17] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4475–4479.

[18] Q. Yu, P. Liu, Z. Wu, S. K. Ang, H. Meng, and L. Cai, "Learning cross-lingual information with multilingual blstm for speech synthesis of low-resource languages," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5545–5549.

[19] H. Zen, N. Braunschweiler, S. Buchholz, M. J. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 6, pp. 1713–1724, 2012.

[20] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Speaker and language factorization in dnn-based tts synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5540–5544.

[21] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for lstm-rnn based statistical parametric speech synthesis," *Interspeech 2016*, pp. 2468–2472, 2016.

[22] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[23] R. Fu, J. Tao, Z. Wen, J. Yi, and T. Wang, "Focusing on attention: Prosody transfer and adaptative optimization strategy for multi-speaker end-to-end speech synthesis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6709–6713.

[24] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *Icassp IEEE International Conference on Acoustics*, 2019.

[25] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[26] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.

[27] E. A. Strickland, "An introduction to the psychology of hearing (6th edition)," *Journal of the Acoustical Society of America*, vol. 136, no. 5, pp. 2898–2899, 2014.

[28] E. Battenberg, R. Skerry-Ryan, S. Mariooryad, D. Stanton, D. Kao, M. Shannon, and T. Bagby, "Location-relative attention mechanisms for robust long-form speech synthesis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6194–6198.

[29] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *ICLR 2018-International Conference on Learning Representations*, 2018.