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

Bi-level Speaker Supervision for One-shot Speech Synthesis

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

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

Abstract

The gap between speaker characteristics of reference speech and synthesized speech remains a challenging problem in one-shot speech synthesis. In this paper, we propose a bi-level speaker supervision framework to close the speaker characteristics gap via supervising the synthesized speech at speaker feature level and speaker identity level. The speaker feature extraction and speaker identity reconstruction are integrated in an end-to-end speech synthesis network, with the one on speaker feature level for closing speaker characteristics and the other on speaker identity level for preserving identity information. This framework guarantees that the synthesized speech has similar speaker characteristics to original speech, and it also ensures the distinguishability between different speakers. Additionally, to solve the influence of speech content on speaker feature extraction task, we propose a text-independent reference encoder (ti-reference encoder) module to extract speaker feature. Experiments on LibriTTS dataset show that our model is able to generate the speech similar to target speaker. Furthermore, we demonstrate that this model can learn meaningful speaker representations by bi-level speaker supervision and ti-reference encoder module.

Index Terms: speech synthesis, one-shot, bi-level speaker supervision, ti-reference encoder

1. Introduction

With the development of deep learning, end-to-end speech synthesis models, such as Tacotron [1] and its varieties [2–4], are proposed to simplify traditional TTS pipeline [5–8] with a single neural network. With the help of WaveNet like neural vocoder [9], the quality and naturalness of synthesized voice are greatly improved. When the corpus of a given speaker is sufficient, the synthesized speech are even comparable with human recordings. When the corpus is only tens of minutes, many speaker adaptive solutions have also been proposed to synthesize speech of target speaker [10–14] and those methods can get effective results.

However, one-shot speech synthesis which using seconds of target speaker's speech to synthesize new speech is still a challenging problem. Due to small amount of data, it is difficult to extract effective speaker features to guide speech synthesis system. This problem results in a gap between the speaker characteristics of synthesized speech and reference speech. To solve the above problem, one-shot speech synthesis training methods can boil down to two aspects. One aspect is training speaker feature extraction task and speech synthesis task separately. In [15], the author uses a speaker extraction module to extract speaker feature, then embeds the speaker feature to guide the multi-speaker speech synthesis model. The speaker extraction module is independently-trained for speaker recognition task. This method can extract effective speaker representation, but it is not optimal for the multi-speaker speech synthesis task and there would be cumulative errors due to segmentation. To overcome the accumulation of errors caused by segmented training, another aspect of one-shot speech synthesis methods is training speaker extraction task and speech synthesis task jointly. In [16], the author uses a style token to transfer target prosody to the synthesized speech. The style token is learned by training tacotron and style encoder network together. In [17], a contrastive triplet loss is used to ensure that speaker embeddings predicted by the identical speaker are closer than the embedding computed from different speakers. This type of method combines speaker feature extraction tasks with multi-speakers speech synthesis tasks together and reduces the accumulation of errors. However, since the speaker representation is learned by unsupervised learning, it is easily affected by speaker accents, speaking content, etc. This may prevent the model from focusing on speaker information, resulting in synthesized speech that is not similar to the reference speaker.

To overcome the above issues, this paper proposes bi-level speaker supervision on the synthesized speech, which combines speaker feature extraction and speaker identity reconstruction to achieve one-shot learning. In order to make the synthesized speech as similar as possible to the target speaker, we supervise the synthesized speech at speaker feature level and speaker identity level respectively. A feature alignment loss and an identity preserving loss are proposed to guarantee the identity consistency of the synthesized speech. To overcome the influence of the different content on the speaker feature extraction, we construct a text-independent reference encoder module. This module takes the speaker identity information and the speaker's arbitrary speech as input, and outputs speaker features that have no relationship to text. In the inference stage, only a few seconds of the target speaker's speech is needed to extract speaker feature. In summary, the main contributions are as follows:

- Bi-level speaker supervision is newly set up for our proposed framework to close the gap between speaker characteristics of synthesized speech and the reference speaker, with the one on speaker feature level contributing to close speaker characteristics and the other on speaker identity level serving for identity maintenance.
- A text-independent reference encoder is proposed to decrease the influence of content on speaker feature extraction. The experiments on LibriTTS dataset show this module can well cluster speech based on speakers.

The rest of the paper is organized as follows. Section 2 presents our proposed framework and ideas in detail. The results and analysis are presented in Section 3. The conclusions and future work are described in Section 4.

http://dx.doi.org/10.21437/Interspeech.2020-1737

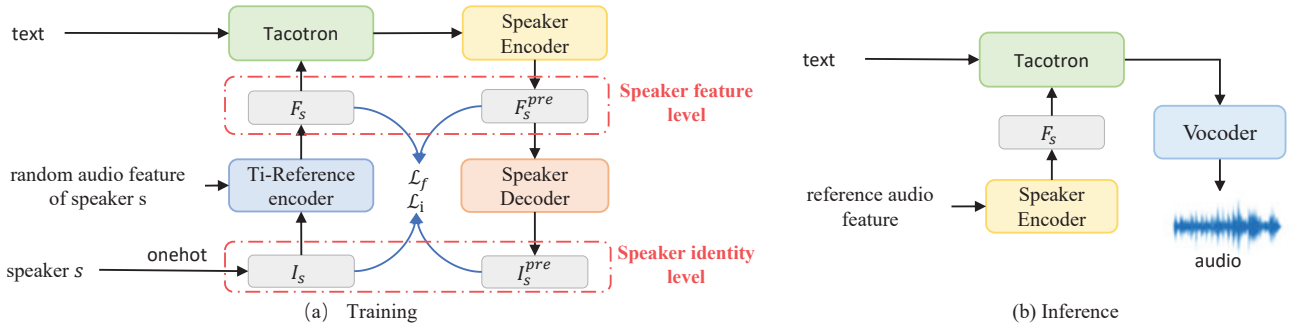

Figure 1: *(a) Training architecture. A one-hot encoded speaker id I_s and the speaker's arbitary acoustic feature are fed to the ti-reference encoder module to calculate speaker feature F_s. F_s is used to condition the tacotron decoder states. The predicted acoustic feature is fed to speaker encoder and speaker decoder to predict the speaker feature F_s^{pre} and speaker id I_s^{pre}. We calculate the loss at speaker feature level and identity level respectively, thereby achieving bi-level speaker supervision. (b) Inference architecture. The reference acoustic feature is fed to speaker encoder to compute speaker feature, and then embedding this feature to tacotron to synthesis speech in reference speaker's voice.*

2. Proposed method

In this section, we will introduce the whole framework of our proposed method firstly, then we will present the ideas about bi-level speaker supervision and the ti-reference encoder.

2.1. Framework

As shown in the left of Fig. 1, in the training procedure, there are four components: a ti-reference encoder, a tacotron model which is an end-to-end speech synthesis system, a speaker encoder and a speaker decoder module. Given a one-hot encoded speaker id and the speaker's reference acoustic feature, the ti-reference encoder will compute the speaker feature, the detail about this module is shown in Subsection 2.3. The tacotron model is used to predict target acoustic feature under the condition of the speaker feature. The speaker encoder and speaker decoder module are used to predict the speaker feature and speaker identity information of the synthesized speech, then we compare them with the input speaker feature and speaker identity.

The inference architecture is shown in the right of Fig. 1. To achieve one-shot mode, we can feed an arbitrary reference acoustic feature to compute speaker feature, then embedding this feature to synthesis speech in reference speaker's voice.

2.2. Bi-level speaker supervision

A good speaker feature space should have two characteristics. One is speaker features of the same speakers should be as similar as possible. The other is speaker features of the different speakers should be distinguishable. In multi-speaker speech synthesis system, there are often two levels of speaker representation. One is the speaker identity level. When training multiple speech synthesis systems, we give each speaker a specific id. This level only emphasizes the differences between different speakers and lacks the similarity of the same speaker. Another is the speaker feature level. In general, the speaker feature of the synthesized speech should be similar to the input speaker feature, but if we only optimize the network through this rule, the speaker feature space can only guarantee the similarity of the same speaker's speaker features, it can not guarantee the difference of different speakers.

Therefore, by supervising the synthesized speech at speaker feature level and speaker identity level, we can learn a good speaker feature space, which can meet the similarity of the same speaker, and also can distinguish the different speakers. In addition, this method can optimize speech synthesis system from the perspective of the speaker and ensure the synthesized speech more similar to the reference speech.

2.2.1. *Speaker feature level supervision*

In the speaker feature's space, the more similar the speaker characteristics are, the closer the distance of the speaker features. A speaker encoder module is used to map the predicted acoustic feature into the speaker feature space to get the speaker feature F_s^{pre}. Compared with the speaker feature F_s which is used to condition the tacotron, the closer the distance between F_s^{pre} and F_s is, the more similar the two speaker characteristics. We propose a speaker feature loss as speaker feature level supervision, which can be expressed as:

$$\mathcal{L}_f = MSE(F_s^{pre}, F_s) \tag{1}$$

2.2.2. *Speaker identity level supervision*

The speaker feature loss $\mathcal{L}_f$ can guarantee similarity of the same speaker's speaker features, but can not guarantee the speaker features' distinction from different speakers. We can reconstruct speaker identity information from speaker features, which illustrates that speaker features have ability to distinguish different speakers. A speaker decoder module is used to decode speaker feature to speaker identity information I_s^{pre}. The goal is that the predicted speaker identity I_s^{pre} should be consistent with the original speaker identity I_s. So, the cross-entropy(CE) between I_s^{pre} and I_s is used as the speaker identity level supervision, which is called identity preserving loss $\mathcal{L}_i$.

$$\mathcal{L}_i = CE(I_s^{pre}, I_s) \tag{2}$$

2.2.3. *Bi-level speaker supervision*

Combining the two constraints above, $\mathcal{L}_f$ and $\mathcal{L}_i$ are adversarial in the training procedure. $\mathcal{L}_f$ guarantee the similarity of the same speaker's speaker features, and $\mathcal{L}_i$ guarantee the distinction of different speakers' speaker features. To preserve the advantages of the two loss functions, we take the weighted sum of $\mathcal{L}_f$ and $\mathcal{L}_i$ as the bi-level speaker supervision loss $\mathcal{L}_f$.

$$\mathcal{L}_{bi} = \lambda_f \mathcal{L}_f + \lambda_i \mathcal{L}_i \tag{3}$$

where λ_f and λ_i are trade-off parameters.

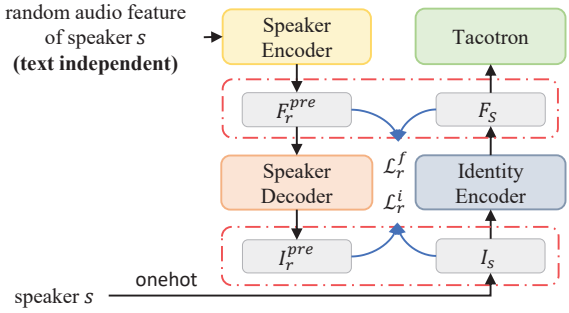

Figure 2: *Ti-reference encoder.*

2.3. Ti-reference encoder

Speaker feature should be independent of the speech's content. The text-independent reference (ti-reference) encoder is based on the ideas of the bi-level speaker supervision. Firstly, an identity encoder module is used to extract speaker features from speaker identity information. The advantage of using speaker identity information to calculate speaker features is that speaker identity information has nothing to do with the content and it is easier to converge when training. Secondly, to further ensure the speaker encoder module is independent of content during the inference phase, we randomly select a piece of acoustic features of input speaker and use speaker encoder and speaker decoder module to calculate the speaker feature and speaker identity information. Subsequently, we can perform bi-level speaker supervision on the input acoustic feature. Because the acoustic features are not related to the text, we can improve the text-independent performance of the speaker encoder module. Besides, it is helpful to ensure that the trained speaker feature space is consistent with the real distribution.

Suppose the speaker feature and speaker identity information calculated from random acoustic feature are expressed as F_r^{pre} and I_r^{pre}, the speaker feature loss $\mathcal{L}_r^f$ and identity preserving loss $\mathcal{L}_r^i$ can be expressed as:

$$\mathcal{L}_r^f = MSE(F_r^{pre}, F_s) \tag{4}$$

$$\mathcal{L}_r^i = CE(I_r^{pre}, I_s) \tag{5}$$

$$\mathcal{L}_{ti} = \lambda_f \mathcal{L}_r^f + \lambda_i \mathcal{L}_r^i \tag{6}$$

$$\mathcal{L}_{total} = \lambda_{bi} \mathcal{L}_{bi} + \lambda_{ti} \mathcal{L}_{ti} \tag{7}$$

$\mathcal{L}_{total}$ is the weighted sum of $\mathcal{L}_{bi}$ in Eq.3 and $\mathcal{L}_{ti}$ in Eq.6 and it participates in optimization of the whole network during training procedure.

3. Experiments

To verify the validity of the method, we train models on the LibriTTS [18]. We mix train-clean-100, train-clean-360 dataset, which has 460 hours and 1146 speakers. We randomly select 5k utterances as the testing set of seen speakers, and the rest utterances are used as training set. The test-clean dataset is used as testing set of unseen speakers and the dev-clean dataset is used as development set. All the wav files are sampled at 24KHz.

3.1. Setup

Acoustic features are extracted with 10 ms window shift. LPCNet [19] is utilized to extract 32-dimensional acoustic features, including 30-dimensional BFCCs, 1-dimensional pitch and 1-dimensional pitch correction parameter.The Tacotron2 framework is the same as in the paper [2]. The structures of other modules proposed in the paper are as follows:

- Speaker encoder module consists of a GRU (128 units for each GRU component), followed by two fully connected networks, in which the dimensions of the output are 128 and 32 respectively.
- Speaker decoder is made up of a three-layer fully connected network, the dimensions from input to output are 256, 512,1160.
- Identity encoder is made up of a three-layer fully connected network. The dimensions from input to output are 512, 256,32.

To explore the effects of bi-level speaker supervision and ti-reference encoder, we perform an ablation study to verify it can help to learn representative speaker feature and can help to improve the similarity of synthesized speech.

- **P** stands for our proposed model which combines bi-level speaker supervision and ti-reference speaker encoder.
- **Ivector** means using ivector as speaker representation to train multi-speakers speech synthesis system.
- **P w/o f** means removing speaker feature level supervision (without $\mathcal{L}_f$) in our proposed model.
- **P w/o i** means removing speaker identity level supervision (without $\mathcal{L}_i$) in our proposed model.
- **P w/o f&i** means removing bi-level speaker supervision (without $\mathcal{L}_f$ and $\mathcal{L}_i$) in our proposed model.
- **P w/o ti** means removing ti-reference encoder model (without $\mathcal{L}_r^f$ and $\mathcal{L}_r^i$) in our proposed model and just uses speaker's id embedding as reference encoder.

3.2. Speaker feature visualization

Firstly, we explore the effect of ti-reference encoder. We randomly select 10 unseen speakers from the test set, then randomly select 200 sentences of each speaker. We use speaker encoder module in the **P** and **P w/o ti** systems to extract speaker features, and visualize those features in 2D space with t-SNE [20] in Fig. 3. We find that the ti-reference encoder is helpful to distinguish different speakers and cluster the sentences of the same speaker. Although only using bi-level supervision on the synthesized speech can learn a certain distinguishing speaker features, when the ti-reference encoder is removed in our proposed model, the effect of clustering will become worse. Since the input acoustic features' content of the ti-reference encoder is arbitrary, the model will focus on the characteristics of speaker and decrease the influence of the speeches' content. To further confirm this assumption, we test the classification accuracy on the seen speakers. We also calculate the distance between the features extracted by identity encoder and speaker encoder modules. This measure can reflect the gap between the speaker features obtained by one-shot and the real. The accuracy and the distance of the feature are shown in Table 1. As we can see, after removed the ti-reference encoder module, the recognition performance will decrease and the distance between the speaker features will increase. The results show that the ti-reference encoder has a significant effect on extracting effective speaker features.

Secondly, we explore the effect of bi-level speaker supervision on the synthesized speech. For the unseen speaker, we

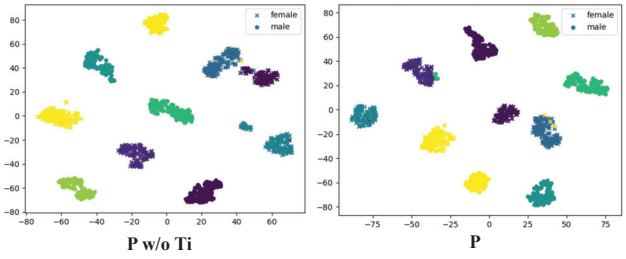

Figure 3: *Visualization of speaker features extracted from unseen speakers's speech. 'x' are female speakers and 'o' are male speakers.*

Table 1: *The ti-reference encoder's effect on speaker features. 'acc' (higher is better) stands the accuracy for speaker identity prediction on speaker features. 'dis' (lower is better) stands the distance between the speakers features extracted by identity encoder and speaker encoder.*

model	acc	dis
P w/o ti	0.716	0.09
P	**0.950**	**0.037**

calculate the input speaker features F_s and the output speaker features F_s^{pre}. We randomly selected 10 unseen speakers. For each speaker , we select 50 speech and 50 synthesized speech accordingly. The input text was fixed. Fig. 4 shows the visualization of speaker features. Compared to the **P w/o f&i**, we can find that the speech synthesized by our proposed model **P** can not only distinguish speaker characteristics well but also be consistent with the original input speaker feature. When there is a lack of feature level supervision, which is the **P w/o f** model, the distance between the original and synthesized speaker features will increase. In the absence of identity level supervision, which is the **P w/o i** model, different speakers can not be distinguished. This also shows that the bi-level speaker supervision can close the gap between speaker characteristics of reference speech and the synthesized speech.

3.3. Naturalness and similarity evaluation

We conduct Mean Opinion Score (MOS) listening test for naturalness and similarity of synthesized speech. 20 listeners participated in the evaluation. In each experimental group, 20 parallel sentences are selected randomly from testing sets of each system. By observing the results of the baseline, our method can get more effective speaker features than baseline. Among all the systems, **P** achieves the best performance. Additionally, we can observe that the bi-level speaker supervision on synthesized speech can improve the similarity performance more than the naturalness.

4. Conclusions

In this paper, we present a multi-speakers speech synthesis training method for one-shot setting which supervising synthesized speech at speaker feature level and speaker identity level. This method can close the gap between speaker characteristics of reference speech and synthesized speech and improve the similarity of the synthesized speech. Specifically, on speaker feature level, speaker feature loss can ensure the speaker feature of synthesized speech are similar to the reference speaker feature. On speaker identity level, identity preserving loss can help synthesized speech retain the identity information of the original speaker. In addition, we propose a ti-reference encoder to reduce the influence of content on speaker feature extraction and obtain more effective speaker features. Experimental results show that through such joint optimization, effective speaker features can be obtained for one-shot speech synthesis, and the synthesized speech has good similarity and naturalness. Further, we will try to explore the physical meaning of each dimension of the learned speaker features, so as to control the attributes of speech from a more detailed perspective.

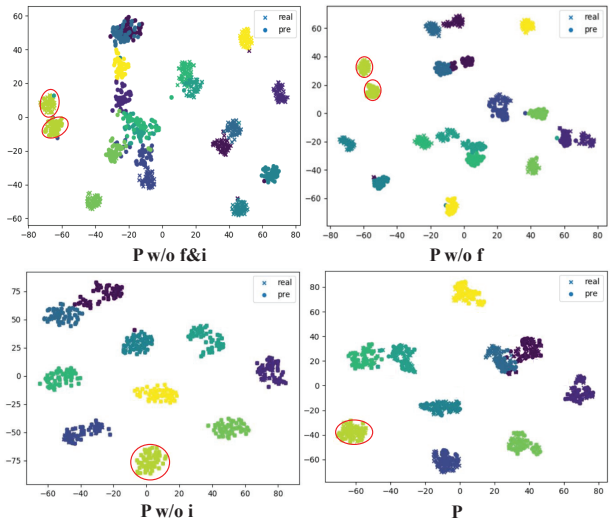

Figure 4: *Visualization of speaker features extracted from unseen speakers' speech and synthesized speech. 'x' are synthesized speech and 'o' are real speech. Real and synthesized speech of the same speaker consistently forms the same cluster in our proposed method (like the example shown in the red circle).*

Table 2: *MOS results for one-shot speech synthesis. Each system evaluates the naturalness and similarity of synthesized speech.*

model	naturalness	similarity
P w/o f&i	3.64	2.79
P w/o f	3.70	2.96
P w/o i	3.68	2.98
P w/o ti	3.56	3.01
Ivector	3.72	3.04
P	**3.98**	**3.20**

5. 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 Inria-CAS Joint Research Project (No.173211KYSB20170061 and No.173211KYSB20190049). This research is (partially) funded by Huawei Noah's Ark Lab. This work is also supported by the CCF-Tencent Open Research Fund.

6. References

[1] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017. [Online]. Available: http://arxiv.org/abs/1703.10135

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 4779–4783.

[3] Y.-A. Chung, Y. Wang, W.-N. Hsu, Y. Zhang, and R. 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.

[4] Y. Wang, R. Skerry-Ryan, Y. Xiao, D. Stanton, J. Shor, E. Battenberg, R. Clark, and R. A. Saurous, "Uncovering latent style factors for expressive speech synthesis," *arXiv preprint arXiv:1711.00520*, 2017.

[5] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[6] A. W. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis." in *in Eurospeech97*, 1997, pp. 601–604.

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

[8] H. Ze, A. Senior, and M. 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.

[9] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[10] 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.

[11] Y. Zhao, D. Saito, and N. Minematsu, "Speaker representations for speaker adaptation in multiple speakers blstm-rnn-based speech synthesis," *space*, vol. 5, no. 6, p. 7, 2016.

[12] P. Swietojanski and S. Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 171–176.

[13] 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.

[14] 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.

[15] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez-Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *CoRR*, vol. abs/1806.04558, 2018. [Online]. Available: http://arxiv.org/abs/1806.04558

[16] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *CoRR*, vol. abs/1803.09047, 2018. [Online]. Available: http://arxiv.org/abs/1803.09047

[17] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," *CoRR*, vol. abs/1802.06984, 2018. [Online]. Available: http://arxiv.org/abs/1802.06984

[18] 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," *ArXiv*, vol. abs/1904.02882, 2019.

[19] J.-M. Valin and J. Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[20] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.