ONE-SHOT VOICE CONVERSION FOR STYLE TRANSFER BASED ON SPEAKER ADAPTATION

Zhichao Wang[1], Qicong Xie[1], Tao Li[1], Hongqiang Du[1], Lei Xie[1], Pengcheng Zhu[2], Mengxiao Bi[2]*

[1]Audio, Speech and Language Processing Group (ASLP@NPU)
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Fuxi AI Lab, NetEase Inc., Hangzhou, China

ABSTRACT

One-shot style transfer is a challenging task, since training on one utterance makes model extremely easy to over-fit to training data and causes low speaker similarity and lack of expressiveness. In this paper, we build on the recognition-synthesis framework and propose a one-shot voice conversion approach for style transfer based on speaker adaptation. First, a speaker normalization module is adopted to remove speaker-related information in bottleneck features extracted by ASR. Second, we adopt weight regularization in the adaptation process to prevent over-fitting caused by using only one utterance from target speaker as training data. Finally, to comprehensively decouple the speech factors, i.e., content, speaker, style, and transfer source style to the target, a prosody module is used to extract prosody representation. Experiments show that our approach is superior to the state-of-the-art one-shot VC systems in terms of style and speaker similarity; additionally, our approach also maintains good speech quality.

Index Terms— voice conversion, one-shot, adaptation, over-fit, style transfer

1. INTRODUCTION

Voice conversion (VC) aims to modify speech from a source speaker to sound like that of a target speaker without changing the linguistic content. Voice conversion based on neural networks, e.g. GAN [1], VAE [2], recognition-synthesis [3] frameworks, have significantly improved speech quality and speaker similarity. Despite recent progress, most current works focus on the transformation of timbre while ignoring the transformation of style contained in source speech. Recently, explicit prosodic features [4, 5] and implicit prosody extractor [6,7] are used to model prosody and perform source style transfer, while requiring a large number of recordings. Data collection is expensive and time-consuming. Therefore, it remains a challenge to build a high-quality voice conversion framework for style transfer with limited data.

Recently, one-shot voice conversion approaches are proposed, which refers to the conversion of the speaker identity given only one utterance of the target speaker, which is more practical for application as compared with previous methods in which sizeable data for the target speaker is needed. Some studies [8–10] propose to extract content and speaker representation from speech during training stage, and concatenate content representation from source speech and speaker representation from target speech at run-time to produce converted speech. However, the content representation may still contain speaker-related information, which results in unstable

performance in terms of speaker similarity. Other works [11, 12] decompose speech into three factors: content, speaker, and pitch with the premise that pitch is highly related to speaker characteristics. These methods further improve the disentanglement ability, while there is still a gap between the converted speech and target speech in terms of speaker similarity. To improve speaker similarity, many works [13–15] in text to speech (TTS), a related task, focus on few-shot speaker adaptation. Unlike few-shot (≥ 10 utterances) TTS, in one-shot voice conversion, training on one utterance (3-4s) can easily lead to over-fitting and reduce the speaker similarity and speech quality. Besides delivering correct content and speaker information, transferring style or prosody from source speaker to the target is also desired, but more challenging [4–7], as prosodic aspects are entangled with content and speaker as well and the over-fit problem is rather complicated.

In our paper, we address the *one-shot style transfer* problem for voice conversion, aiming to transferring the source content and style to the target speaker while maintaining the good target speaker identity; more challengingly, we achieve the above goal with only one utterance with several seconds from the target speaker. Specifically, to solve the over-fit problem caused by training with only one utterance, we propose a novel one-shot voice conversion framework for style transfer based on the recognition-synthesis framework [3], integrating speaker normalization [16], weight regularization [17], and prosody modeling. Specifically, speaker normalization is used to remove speaker-related information in the bottleneck feature (BN) extracted from ASR, which is helpful to improve speaker similarity. Weight regularization is applied to the adaptation with one utterance to prevent performance degradation. A prosody module [7] is adopted to explicitly extract prosody information for transfer. We believe such a comprehensive characterization of speech factors can achieve good style transfer and make the model focus on learning timbre in the adaptation process. The experimental results on VCTK [18] and CMU-ARCTIC [19] show that over-fitting is greatly alleviated after using these methods, and our framework is better than several state-of-the-art (SOTA) methods including AGAINVC [9], GSE [11] and VQMIVC [12] in terms of style and speaker similarity.

The rest of the paper is organized as follows. Section 2 introduces our proposed methods. Section 3 presents the experiments and compares our proposed system with SOTA systems with subjective and objective evaluation. Section 5 concludes this paper.

2. PROPOSED APPROACH

2.1. System Overview

Our proposed one-shot voice conversion framework is based on a recognition-synthesis architecture, which is shown in Fig. 1. This
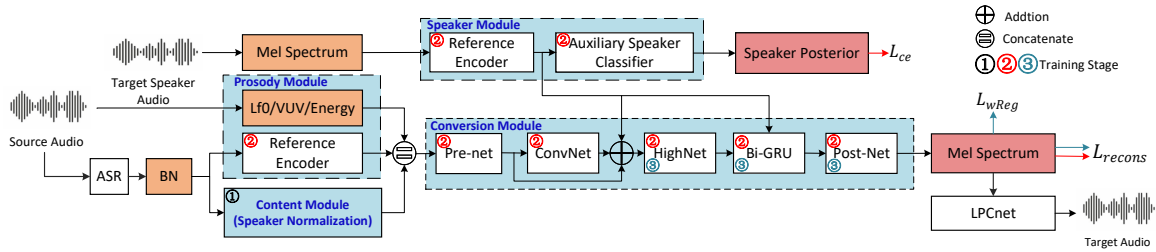

Fig. 1. Overview of the components of the proposed system.

framework consists of four main modules (light blue boxes in Fig. 1): content module, speaker module, prosody module and conversion module. The content module takes bottleneck features (BN) extracted from ASR as input, and the output is speaker-independent content representation. Specifically, we adopt speaker normalization technique to remove speaker-related information contained in BN. The speaker module takes mel spectrum from target speaker to extract the speaker representation. To ensure the speaker module is able to extract desired speaker representation without speaker confusion, we add an auxiliary speaker classifier after the reference encoder. The prosody module learns to extract speaker-independent prosody representations, which integrates explicit and implicit hybrid modeling methods [7]. The explicit modeling includes raw logarithmic domain fundamental frequency (lf0), normalized short-term average amplitude (energy), and the voice/unvoice flag (VUV). For implicit modeling, we adopt reference encoder to extract prosody representation. The conversion module takes prosody, content, and speaker representations as input, and the output is mel spectrum. Finally, modified LPCnet [20] is adopted to reconstruct waveform from mel spectrum.

2.2. Speaker Normalization

Although an ASR system is trained using multi-speaker data, targeting to speaker-independent acoustic representation, studies [7, 21, 22] show that the bottleneck feature (BN) still inevitably contains speaker-related information, such as timbre and style. Thus for voice conversion, the speaker similarity of the converted speech may be degraded. As shown in Fig. 1, we specifically introduce a speaker normalization method to remove the speaker-related information. Previously, adversarial training [21, 22] was usually used to achieve this goal. But taking into account the training cost and instability of adversarial training, we utilize a speaker normalization method [16] instead in our framework. Specifically, we utilize an any-to-one VC network here to serve as a speaker 'normalization' trick to 'normalize' the source audio's BN to mel spectrum of a specific speaker to achieve the purpose of normalizing content information. By doing so, we actually relieve the burden of the rest of our framework.

2.3. Weight Regularization

As shown in Fig. 1, in order to alleviate the over-fitting problem caused by training with only one sentence and improve the stability of the training process, we introduce weight regularization [17]. The weight regularization is a variant of l2 regularization, defined as

$$L_{wReg} = ||\theta - \theta_f||^2, \tag{1}$$

where θ represents the parameters of HighNet, Bi-GRU and post-net updated over time in adaptation process, and θ_f represents these three layers' parameters before adaptation which is used as fixed value during the adaptation process. The core idea in L_{wReg} is to prevent the parameters of adapted model to drift far away from those of the base model learnt in the large dataset. The regularization prevents model from changing too significantly, which is helpful in improving the stability of the training process.

2.4. Source Style Transfer

Many one-shot voice conversion methods [8, 9, 11, 12, 23] focus on disentangling content and speaker, while ignoring the prosody modeling, which results in that the prosody information may leak to the content and speaker representation. This reduces the stability of disentanglement and causes poor speaker similarity, although the prosody of source is much transferred to the target. Inspired by our previous work [7], we add a prosody module to the one-shot voice conversion framework. Recall that we explicitly decompose speech into three parts: content, speaker as well as style, and this prosody module is specifically in charge of the style representation for the final conversion module. In detail, the prosody module includes explicit and implicit modeling schemes. For explicit modeling, we extract energy, VUV, and raw lf0 from source audio. For implicit modeling, the BN is used as the input of a reference encoder, and the output is prosody representation. By adding the prosody module, the voice conversion framework is able to perform style transfer.

2.5. Training and Conversion Procedure

As shown in Fig. 1, our approach is composed of three training phases and a conversion phase. Different phases are marked with numbers in the figure.

Training phase 1. As we mentioned in Section 2.3, we utilize an any-to-one VC method to implement our content module. The content module is trained on a large amount of speech data and fine-tunes on a specific speaker to ensure the performance of speaker normalization. In our experiment, the content module is optimized with mel reconstruction loss.

Training phase 2. The whole model is trained with a large amount of data. Note that the speaker module is trained with a randomly selected mel spectrum of the same speaker to make sure that the extracted speaker embedding is only related to the speaker information. The prosody and conversion modules are optimized with reconstruction loss L_{recons}. The speaker module is optimized with cross-entropy loss L_{ce}. The loss function of this stage can be described as follows:

$$Loss_{stage2} = L_{recons} + L_{ce} \tag{2}$$

Table 1. MOS results with 95% confidence interval.

	Model		Speech Quality	Style Similarity	Speaker Similarity
Comparison	AGAINVC [9]		2.81±0.12	3.03±0.34	2.96±0.34
	VQMIVC [12]		2.94±0.27	3.19±0.30	3.02±0.15
	GSE [11]		**3.46±0.16**	3.19±0.12	3.03±0.18
	GSE-finetune		3.09±0.14	2.88±0.11	3.31±0.12
Ablation	BL	Baseline	2.93±0.13	2.84±0.16	3.25±0.13
	P1	+Speaker Normalization	3.16±0.12	3.15±0.12	3.37±0.10
	P2	+Weight Regularization	3.27±0.12	3.21±0.13	3.43±0.07
P3 (Proposed)		+Prosody Module	3.43±0.12	**3.76±0.17**	**3.56±0.15**

Training phase 3. This stage is for adaptation. In this stage, we only use one utterance from target speaker to fine-tune the model. Previous works [13, 15] confirm that a large number of parameters may make model easy to over-fit. Hence, only part of the conversion module is involved in training, including HighNet, Bi-GRU and Post-Net, as shown in Fig. 1. To further prevent performance degradation caused by over-fitting, we adopt the weight regularization [17] in this phase. The loss used in this phase is described as follows:

$$Loss_{stage3} = L_{recons} + \gamma L_{wReg} \qquad (3)$$

Conversion phase. The content and prosody representation are extracted from source speech. Note that we use linear transformation to obtain converted f0. The speaker representation is extracted from target speech. The conversion module takes content, prosody, and speaker representation to reconstruct converted speech. Through the proposed disentanglement and training procedure, the converted speech is expected to have the same prosody as source speech, while maintaining the content and target speaker's identity.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and Experimental Setup

In our experiments, 102 speakers from VCTK [18] are used to train the conversion model. For one-shot testing, *p340, p363* from VCTK, as well as *slt*, *bdl* from CMU-ARCTIC [19] are used as the target speakers. The duration of target speech ranges from 3 to 4 seconds. For the source audio, we use CMU-ARCTIC (rms, clb) and ESD dataset [24]. All speech utterances are downsampled to 16kHz. We use 80-dim mel spectrum computed with 50ms frame length and 12.5ms frame shift. The ASR system is a TDNN-F model trained with LibriSpeech [25] corpus containing 1k hours speech. We use the 256-dim bottleneck features as the linguistic representation, which is extracted from the last fully-connected layer before softmax in the TDNN-F model. Modified LPCnet [20] is adopted to reconstruct waveform from mel spectrum. The vocoder is trained with speech data of 102 speakers from VCTK.

To validate our proposed method, we implement comparison and ablation systems. For comparison systems, we selected three SOTA one-shot VC methods, including AGAINVC [9], GSE [11] and VQMIVC [12]. For a fair comparison, we finetune GSE with target speaker's utterance, which is referred as GSE-finetune. For ablation analysis, we implement four systems BL, P1, P2, and P3 (proposed). Among them, BL is composed of speaker and conversion module to convert BN to mel spectrum. P1 uses speaker normalization based on BL, and P2 adopts weight regularization based

on P1. P3 is our final proposed system combining all the contributions.

For our proposed method, the content adopts the same model configuration as Tian *et al.* [26]. The content module follows a typical encoder-decoder architecture using CBHG as the encoder and an auto-regressive module consisting of prenet, decoder RNN and postnet as the decoder. For the conversion module, the configuration of CBHG is the same as that of the content module. Prenet consists of 2 fully connected layers with 80 and 256 hidden units respectively. Postnet contains 4 1D-convolution layers with 3*3 kernel size and 256 filters and a fully connected layers with 80 hidden units. The architecture and hyper-parameters of the reference encoder follow the original configuration in [27]. This structure mainly includes 6 layers of convolution and GRU. Each convolution layer is composed of 3×3 filters with 2×2 stride, SAME padding, and ReLU activation. The number of filters in each layer is [32, 32, 64, 64, 128, 128]. Batch normalization is applied to every layer. The output of convolution layers is fed into GRU with 128 units. The speaker classifier consists of 3 fully connected layers. In training stage 2, the conversion model is trained for 120 epochs using batch size of 16. We use Adam optimizer to optimize our model with learning rate decay, which starts from 0.001 and decays every 20 epochs with decay rate 0.7. In training stage 3, we train the conversion model for 2000 steps using one utterance, and γ is set to 1. The learning rate starts from 0.001 and decays every 200 steps in decay rate 0.5.

3.2. Subjective evaluation

We conduct the following listening tests: three mean opinion score (MOS) tests to assess speech quality, style similarity, and speaker similarity respectively. Style similarity measures styles between converted audio and source audio. We select three sentences from rms, clb, and ESD respectively as source audio. Nine sentences are converted to four target speakers (p340, p363, slt, bdl), a total of 36 sentences are used for listening test. We highly recommend readers to listen to our samples*.

Comparison analysis. We compare the proposed method with SOTA one-shot VC methods. The results of MOS tests are shown in Table 1 comparison part. It is observed that GSE achieves the best result in terms of speech quality, and our proposed system P3 gets significantly higher MOS scores in terms of style and speaker similarity than the comparison systems. Previous one-shot methods lack prosody modeling ability, and target speaker's timber is unknown to model, which leads to low speaker similarity and unstable performance. Comparing GSE with GSE-finetune, we can see that making

---

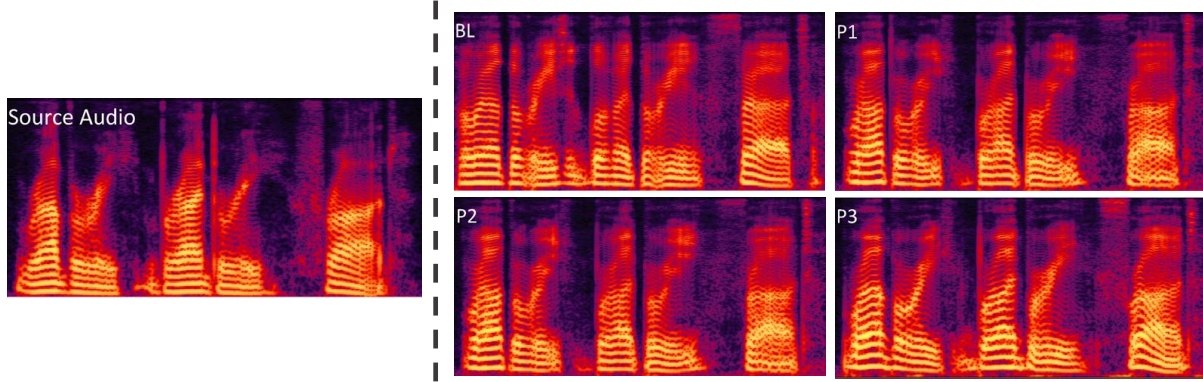*Samples can be found in `https://kerwinchao.github.io/Oneshotvc.github.io/`

Fig. 2. Spectrograms for a source audio utterance (left) and its corresponding converted utterances by different systems (right). The formants from the proposed system (esp. P3) are more similar to the source, indicating good style transfer. By contrast, the formants from BL are too flat.

GSE learn from the target speaker utterance can significantly improve the speaker similarity.

Ablation analysis. As shown in Table 1 ablation part, we evaluated the ablation systems. The BL system obtains poor results in three MOS tests, which indicates that BL system is easy to over-fit. By using our proposed method, the phenomenon of over-fitting is alleviated, and all MOS scores are improved. Adding the prosody module makes the speaker module focus on extracting timber from only one utterance, therefore that style and speaker similarity are able to be improved.

Varying duration. We further evaluate the performance of the proposed system under different duration of utterances (1, 3, 6, 9, 15 seconds) from CMU-ARCTIC speakers. Note that here only CMU-ARCTIC speakers are used for subjective test as listeners need to assess a large set of audio samples. As shown in Fig. 3, the result is affected by the duration of the target speech in general. For the extreme cases, e.g. 1-3s, the model still shows quality degradation caused by over-fitting even if our proposed method is used. But from 3 to 6s, benefiting from the proposed approach, the three MOS scores have clear increased while speech quality and style similarity undergo a quick boost. After 6 seconds, all three curves do not improve significantly and begin to stabilize from 9 seconds.
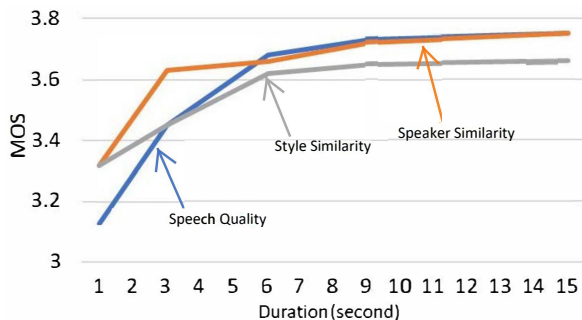

Fig. 3. The MOS performance of proposed system vs. training utterance with different duration

3.3. Objective evaluation

Over-fitting on spectrograms. We visualize the spectrogram of the converted speech for further investigation. The spectrogram of a testing sample is shown in Figure 2. We can see that BL has a very flat formant on the spectrogram, and noise appears in the silence part. Comparing these spectrograms, we observe that over-fitting phenomenon reduces the frequency of the formant and the fundamental frequency, which affects speech quality and speaker similarity. These figures suggest that the details of the frequency and the fluctuation of the formant become more abundant when the proposed method is used. This experiment shows that our method is effective in overcoming over-fitting.

Prosody correlation. To further verify the statistical significance of the expressiveness of each system, we extracted features related to the prosody: frame-level energy and lf0. We use 36 utterances to calculate the Pearson correlation coefficients between source audio and converted audio. The higher the Pearson correlation coefficient of the model, the higher the accuracy of the predicted prosodic attributes. As shown in Table 2, P3 gets the highest scores from the perspective of energy and lf0. This indicates that the prosody module can improve the performance of style transfer. The conclusion from objective measurement is line with the that from subjective listening.

Table 2. Pearson correlation in energy and lf0.

	BL	P1	P2	P3
Energy	0.765	0.727	0.729	**0.755**
Lf0	0.544	0.518	0.531	**0.733**

4. CONCLUSION

In this study, we propose a novel one-shot voice conversion framework for style transfer. This task is challenging as training on one utterance is easy to over-fit, which results in serious degradation of speaker similarity and style. To mitigate this challenge, we build on the recognition-synthesis framework and introduce a disentangled structure to explicitly model content, speaker identity and prosody which are originally entangled in speech. Specifically, we first adopt speaker normalization in content module to normalize speaker-related information. Furthermore, we add weight regularization during adaptation process to prevent over-fitting. Finally, to improve the expressiveness of converted speech, prosody module is added to one-shot voice conversion framework, which can extract rich prosody representation from source audio. Experimental results show that our proposed system outperforms several SOTA one-shot systems in terms of speaker similarity and style.

5. REFERENCES

[1] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice Conversion from Unaligned Corpora Using Variational Autoencoding Wasserstein Generative Adversarial Networks," in *Proc. Interspeech*, 2017, pp. 3364–3368.

[2] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Y. Tsao, and H. Wang, "Voice conversion from Non-Parallel Corpora Using Variational Auto-Encoder," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016, pp. 1–6.

[3] Lifa Sun, K. Li, Hao Wang, Shiyin Kang, and H. Meng, "Phonetic Posteriorgrams for Many-to-One Voice Conversion without Parallel Data Training," in *International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[4] Huaiping Ming, D. Huang, Lei Xie, Shaofei Zhang, Minghui Dong, and Haizhou Li, "Exemplar-based Sparse Representation of Timbre and Prosody for Voice Conversion," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5175–5179.

[5] Tuomo Raitio, Ramya Rasipuram, and Dan Castellani, "Controllable Neural Text-to-Speech Synthesis Using Intuitive Prosodic Features," in *Proc. Interspeech*, 2020, pp. 4432–4436.

[6] Songxiang Liu, Yuewen Cao, Shiyin Kang, Na Hu, Xunying Liu, Dan Su, Dong Yu, and Helen Meng, "Transferring Source Style in Non-Parallel Voice Conversion," in *Proc. Interspeech*, 2020, pp. 4721–4725.

[7] Zhichao Wang, Xinyong Zhou, Fengyu Yang, Tao Li, Hongqiang Du, Lei Xie, Wendong Gan, Haitao Chen, and Hai Li, "Enriching Source Style Transfer in Recognition-Synthesis Based Non-Parallel Voice Conversion," in *Proc. Interspeech*, 2021, pp. 831–835.

[8] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, "Autovc: Zero-shot Voice Style Transfer with Only Autoencoder Loss," in *International Conference on Machine Learning (ICML)*, 2019, pp. 5210–5219.

[9] Yen-Hao Chen, Da-Yi Wu, Tsung-Han Wu, and Hung-yi Lee, "Again-VC: A One-Shot Voice Conversion Using Activation Guidance and Adaptive Instance Normalization," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5954–5958.

[10] Hongqiang Du and Lei Xie, "Improving Robustness of One-Shot Voice Conversion with Deep Discriminative Speaker Encoder," in *Proc. Interspeech*, 2021, pp. 1379–1383.

[11] Hui Lu, Zhiyong Wu, Dongyang Dai, Runnan Li, Shiyin Kang, Jia Jia, and Helen Meng, "One-Shot Voice Conversion with Global Speaker Embeddings," in *Proc. Interspeech*, 2019, pp. 669–673.

[12] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng, "VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-Shot Voice Conversion," in *Proc. Interspeech*, 2021, pp. 1344–1348.

[13] Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Tie-Yan Liu, et al., "AdaSpeech: Adaptive Text to Speech for Custom Voice," in *International Conference on Learning Representations (ICLR)*, 2020.

[14] Sercan Ö. Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou, "Neural Voice Cloning with a Few Samples," in *NeurIPS*, 2018.

[15] Ji-Hoon Kim, Sang-Hoon Lee, Ji-Hyun Lee, Hong-Gyu Jung, and Seong-Whan Lee, "GC-TTS: Few-shot Speaker Adaptation with Geometric Constraints," in *International Conference on Systems, Man, and Cybernetics (SMC)*, 2021.

[16] Thomas Glarner, Janek Ebbers, and Reinhold Häb-Umbach, "Voice Conversion Based Speaker Normalization for Acoustic Unit Discovery," *ArXiv*, vol. abs/2105.01786, 2021.

[17] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu, "Model Adaptation: Unsupervised Domain Adaptation without Source Data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9641–9650.

[18] Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," 2016, University of Edinburgh. The Centre for Speech Technology Research (CSTR).

[19] John Kominek and Alan W. Black, "The CMU Arctic Speech Databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[20] Jean-Marc Valin and Jan Skoglund, "LPCNet: Improving Neural Speech Synthesis Through Linear Prediction," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5891–5895.

[21] Shaojin Ding, Guanlong Zhao, and Ricardo Gutierrez-Osuna, "Improving the Speaker Identity of Non-Parallel Many-to-Many Voice Conversion with Adversarial Speaker Recognition," in *Proc. Interspeech*, 2020, pp. 776–780.

[22] Zhichao Wang, Wenshuo Ge, Xiong Wang, Shan Yang, Wendong Gan, H. Chen, Hai Li, L. Xie, and Xiulin Li, "Accent and Speaker Disentanglement in Many-to-many Voice Conversion," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, pp. 1–5.

[23] Ju chieh Chou and Hung-Yi Lee, "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization," in *Proc. Interspeech*, 2019, pp. 664–668.

[24] Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li, "Seen and Unseen Emotional Style Transfer for Voice Conversion with A New Emotional Speech Dataset," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 920–924, 2021.

[25] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR Corpus Based On Public Domain Audio Books," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[26] Xiaohai Tian, Zhichao Wang, Shan Yang, Xinyong Zhou, Hongqiang Du, Yi Zhou, Mingyang Zhang, Kun Zhou, Berrak Sisman, Lei Xie, and Haizhou Li, "The NUS & NWPU System for Voice Conversion Challenge 2020," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 170–174.

[27] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous, "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," in *International Conference on Machine Learning (ICML)*, 2018, pp. 4693–4702.