VC-AUG : Voice Conversion based Data Augmentation for Text-Dependent Speaker Verification

Xiaoyi Qin[1], Yaogen Yang[1], Yao Shi[1], Lin Yang[2], Xuyang Wang[2], Junjie Wang[2], and Ming Li[1]

[1] Data Science Research Center, Duke Kunshan University, Kunshan, China
{xiaoyi.qin,yaogen.yang,yao.shi,ming.li369}@dukekunshan.edu.cn
[2] AI Lab of Lenovo Research, Beijing, China
{yanglin13,wangxy60,wangjj9}@lenovo.com

Abstract. In this paper, we focus on improving the performance of the text-dependent speaker verification system in the scenario of limited training data. The deep learning based text-dependent speaker verification system generally needs a large-scale text-dependent training data set which could be both labor and cost expensive, especially for customized new wake-up words. In recent studies, voice conversion systems that can generate high quality synthesized speech of seen and unseen speakers have been proposed. Inspired by those works, we adopt two different voice conversion methods as well as the very simple re-sampling approach to generate new text-dependent speech samples for data augmentation purposes. Experimental results show that the proposed method significantly improves the Equal Error Rate performance from 6.51% to 4.48% in the scenario of limited training data. In addition, we also explore the out-of-set and unseen speaker voice conversion based data augmentation.

Keywords: speaker verification · voices conversion · text-dependent · data augmentation

1 Introduction

Speaker verification technology aims to determine whether the test utterance is indeed spoken by the enrollment speaker. In recent years, x-vectors [23] demonstrate state-of-the-art results in the speaker verification field. Multiple different backbone architectures, e.g. TDNN [23], ResNet [2], and their variants [18], etc. are proposed for the front-end feature extraction.

Futhermore, the research works of deep learning based speaker verification also enjoy those publicly open and free speech databases, e.g., AISHELL2 [7], Librispeech [17], Voxceleb1&2 [16, 4] in the text-independent field, and RSR2015 [15], HIMIA [19], MobvoiHotwords in the text-dependent field, etc. Methods in [10, 24] achieve a good performance in the text-dependent speaker verification task if a large amount of text-dependent training data are available. However, it is both labor expensive and time consuming to collect the database. With the

rise of smart home and Internet of Things applications, there are great demands for text-dependent speaker verification, with customized wake-up words. It is almost impossible to collect the corresponding text-dependent speech data for each customized wake-up word.

In recent studies, the speech signals generated by the multi-speaker Text-to-Speech (TTS) and one-to-many or many-to-many voice conversion (VC) systems are getting harder to be distinguished between real-person voice and synthesized voice [29, 27, 5]. So, it is natural to adopt TTS or VC as a data augmentation strategy for speaker verification under the limited training data scenario [12]. The multi-speaker TTS system could create a large amount of speech data from multiple target speakers with different lexical contents. However, in the context of text-dependent cases, since the input text is the same, the synthesized speech data are very similar even for different target speakers. Moreover, different from multi-speaker TTS, the VC system can generate data with various kinds of styles all with the same text-dependent content. Therefore, the VC approaches are more appropriate than TTS as the data augmentation method for text-dependent speaker verification.

This paper aims to improve the text-dependent speaker verification system's performance with a limited number of speakers and training data.

- Limited training data for each speaker. The number of text-dependent utterances of each speaker is less than 10.
- Limited speakers for training. The number of speakers is less than 500.

Targeting the aforementioned scenarios, we propose to train a voice conversion model with limited existing text-dependent data to generate more new text-dependent data. We use two different voice conversion methods as our data augmentation systems. The first one is a Mel-to-Mel voice conversion system [26] using the conditional Seq-to-Seq neural network framework with dual speaker embeddings as the inputs while the other one is a PPP-to-Mel system that converts the phoneme posterior probability(PPP) features [11] with target speaker embedding into Mel-spectrograms [28]. Furthermore, in the limited speaker number case, we adopt the pitch shift(speed perturbation with re-sampling) strategy to augment more speakers. Besides, we also attempt to use the out-of-set unseen speakers' embeddings to generate the text-dependent data from out-of-set speakers. In order to compare TTS and VC based data augmentation methods in the text-dependent speaker verification task, we also train a popular one-hot multi-speaker TTS framework. The ResNet34-GSP [2] model is adopted as the speaker verification system to evaluate different systems.

The paper is organized as follows. Section 2 describes the related works about voice conversion and speaker verification we adopted in this paper. The proposed methods and strategies are described in section 3. Section 4 shows the experimental results. Finally, the conclusion is provided in section 5.

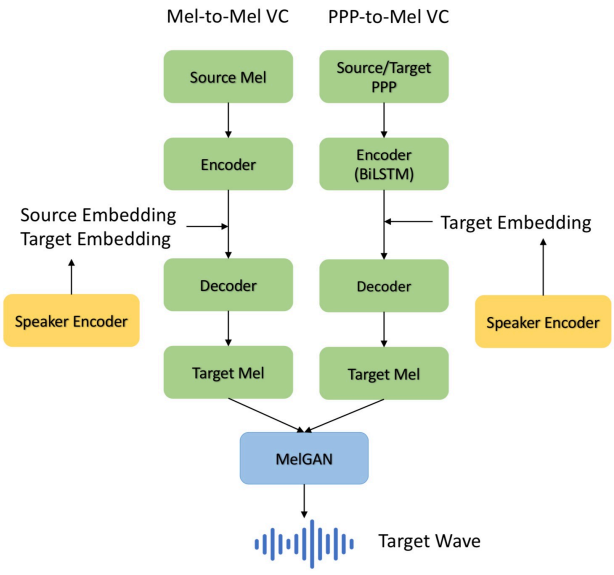

Fig. 1. The architectures of two voice conversion systems used in this work

2 Related Works

2.1 Speaker Verification System

In this paper, we adopt the same structure as [2]. The network structure contains three main components: a front-end pattern extractor, an encoder layer, and a back-end classifier. The ResNet34 [9] structure is employed as the front-end pattern extractor, which learns a frame-level representation from the input acoustic feature. The global statistic pooling (GSP) layer, which computes the mean and standard deviation of the output feature maps, can project the variable length input to the fixed-length vector. The output of a fully connected layer following after the pooling layer is adopted as the speaker embedding layer. The ArcFace [6] (s=32,m=0.2) which could increase intra-speaker distances while ensuring inter-speaker compactness is used as a classifier . The detailed configuration of the neural network is the same with [21]. The cosine similarity serves as the back-end scoring method.

2.2 Voice Conversion System

Mel-to-Mel VC System Firstly, we introduce a many-to-many voice conversion model using the conditional sequence-to-sequence neural network framework with dual speaker embedding [26]. The model is trained on many different source-target speaker pairs, which requires the speaker embeddings from both the source speaker and the target speaker as the auxiliary inputs. To improve

speaker similarity between reference speech and converted speech, we use a feedback constraint mechanism [3], which adds an auxiliary speaker identity loss in the network. This model is named as the Mel-to-Mel VC system because the model directly maps the source speaker Mel-spectrogram to target speaker Mel-spectrogram.

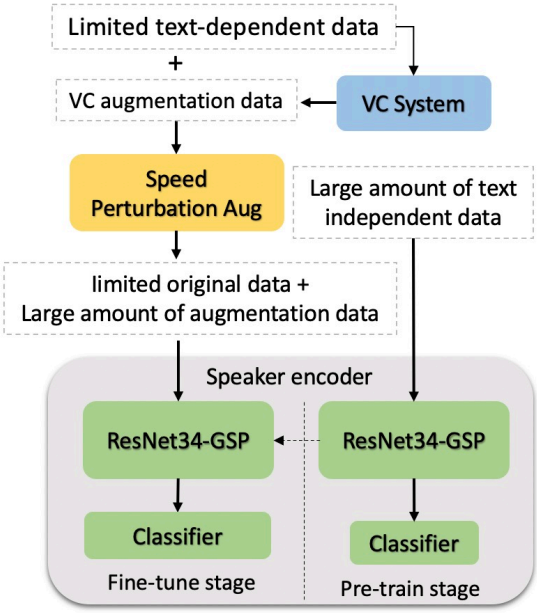

Fig. 2. The pipeline of Data Augmentation based on Voice Conversion in Text-Dependent Speaker Verification.

PPP-to-Mel VC System Besides, we also introduced another VC system. The model is proposed in [28]. First, we use a DNN based auto-speech recognition (ASR) acoustic model, trained on the AISHELL-2 database, to obtain the target speaker phoneme posterior probabilities(PPP) features as the voice conversion

Table 1. The dataset usage of training VC and TTS systems.

Model	Dataset.	Training Spk/Utt Num
ASV	HIMIA	340/3060
VC (PPP-to-Mel)	HIMIA	340/3060
VC (Mel-to-Mel)	HIMIA	340/3060
TTS	DIDI-speech	500/53425

model's input. The model's output is the target Mel-spectrogram feature. In the testing, the source PPP will be assumed to be exactly the same as the target PPP to generate the results. This system is named as the PPP-to-Mel VC system in this paper. The PPP-to-Mel VC system architecture is similar to the Mel-to-Mel VC system expect that there is no feedback constraint. Besides, since the system's input is the target speaker feature rather than the source speaker feature, the input PPP feature is selected randomly from the limited training data.

Fig.1 shows the architectures of two voice conversion systems. The speaker encoder component is the same as the aforementioned ResNet34-GSP model. The vocoder MelGAN [14] is used to reconstruct the time-domain waveform from the predicted Mel-spectrogram.

3 Methods

In this section, the VC data augmentation strategies and the speed perturbation method are introduced in detail. The pipeline of our proposed data augmentation strategy is shown in Fig.2. Those methods are all focused on the limited text-dependent data scenario. In this experiment, we adopt the HIMIA database with 340 speakers [19]. 9 utterances of each speaker in the HIMIA database are randomly chosen as the limited text-dependent data to train the baseline system. Therefore, only 3060 utterances (total have 340*9=3060 utterances) are used to train the VC conversion and fine-tune speaker verification models. The close-talk text-dependent data of the FFSVC20 challenge [21] are chosen as the test data. The trial file can be download from trial_file[3]. Since 3060 sentences with only 'ni hao,mi ya' text are not enough to train a TTS system, we use DiDi-speech [8] with 500 speaker to train a multi-speaker TTS system. The dataset usage of training VC and TTS system show in the Table.1

3.1 Pre-training and Fine-tuning

According to our previous works[20, 19], fine-tuning is an effective transfer learning approach to improve the speaker verification system performance in the limited training data scenario. In this work, we pre-trained the deep speaker verification network with a large-scale text-independent mix-dataset. There are in total 3742 speakers in the pre-training dataset, including AISHELL-2 [7], SLR68[4] and SLR62[5] from openslr.org. These three databases are also considered as out-of-set unseen speaker data for the VC augmentation system. The model was trained for 200 epochs in the pre-training stage, with an initial learning rate of 0.1. The network was optimized by stochastic gradient descent(SGD). All weights in the network remain trainable with an initial learning rate of 0.01 during the fine-tuning stage.

[3] https://github.com/qinxiaoyi/VCaug_ASV
[4] https://openslr.org/68/
[5] https://openslr.org/62/

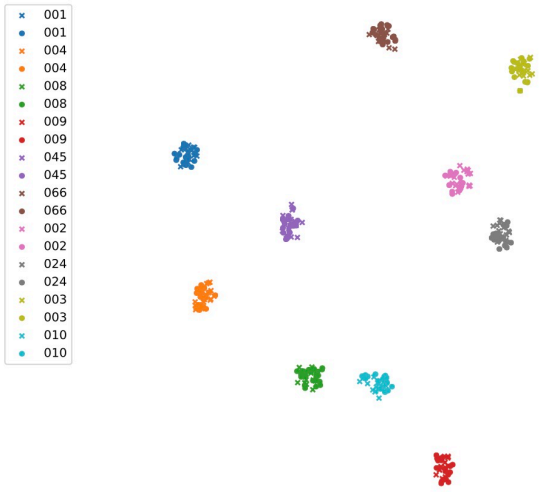

Fig. 3. Speaker embedding visualization by t-SNE for the in-set data. The voice conversion data is generate by PPP-to-Mel VC system. The $*$ stands for the original data and the $\bullet$ stands for voice conversion data

3.2 Data augmentation based on the VC system

The training data of both VC systems is only 3060 utterances with 'ni hao, mi ya' text.

Data augmentation using the Mel-to-Mel VC system For training the Mel-to-Mel VC system, the loss function of the many-to-many voice conversion model is

$$\begin{aligned}\mathcal{L}_{total} = \mathcal{L}_{mel\_before} + \mathcal{L}_{mel\_after} + \mathcal{L}_{stop\_token\_loss} \\ + 5 * \mathcal{L}_{embedding\_loss} + \mathcal{L}_{regular\_loss}\end{aligned} \tag{1}$$

The loss function is also described in detail in[3]. To make the speaker embedding of the voice generated by the voice conversion model close to the target speaker embedding, we increased the weight of embedding loss and set it to 5.

After that, we generated 200 utterances for each target speaker based on a trained Mel-to-Mel VC system. For every target speaker, the source speech of VC's input was random chosen from the other 339 speaker utterances. The embeddings generated by the VC system were computed the cosine similarity with target speaker embedding to handle the outlier. The data with similarity greater than 0.6 are retained.

The limited text-dependent training data (3060 utts) are adopted as source speech for the out-of-set unseen speaker augmentation. Each out-of-set unseen speaker has 20 VC generated text-dependent utterances. After that, the generated data with cosine similarity less than 0.3 are filtered out. Since the out-of-set

Table 2. The performance of the text dependent speaker verification systems under different data augmentation methods. the 9*utt* denotes the limited training data scenario, each speaker only has 9 utterances; the VC AUG$_{in}$ and the VC AUG$_{out}$ denotes the voice conversion data from in-set and out-of-set speakers respectively; the Pitch shift AUG denotes the SoX *speed* function based pitch shift augmentation method.

Model	Training data	Spk / Utt Num.	EER[%]	mDCF$_{0.1}$
Pre-train model	AISHELL2 +SLR62 +SLR68	3472 / 518864	6.51	0.265
Fine-tune model	9 Utts per spk (baseline system)	340 / 3060	7.63	0.331
	+ Pitch shift AUG	1020 / 9180	5.76	0.248
	+ VC AUG$_{in}$ (Mel-to-Mel)	340 / 26160	6.36	0.304
	+ VC AUG$_{in}$ (PPP-to-Mel)	340 / 29089	5.16	0.249
	+ VC AUG$_{out}$ (Mel-to-Mel)	3210 / 48890	6.08	0.295
	+ VC AUG$_{in}$ (Mel-to-Mel) + Pitch shift AUG	1020 / 76978	5.19	0.241
	+ VC AUG$_{in}$(PPP-to-Mel) + Pitch shift AUG	1020 / 87267	**4.48**	**0.212**
	+ TTS (DiDi)	792 / 7323	6.01	0.292

voice conversion is a challenging task, the threshold is not very strict (the most out of set embedding similarity is less than 0.5).

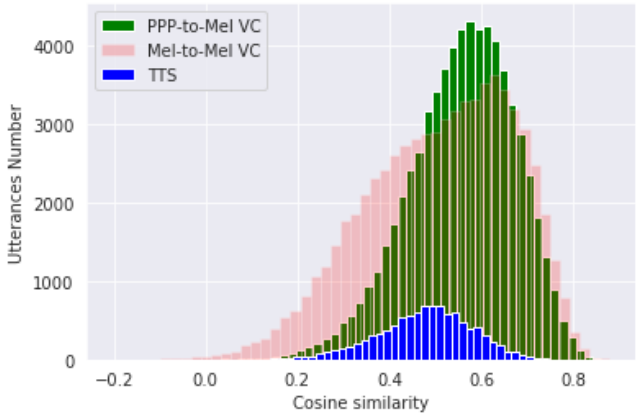

Fig. 4. Histogram of cosine similarity score on the in-set experiment.

Data augmentation using the PPP-to-Mel VC system The procedure the PPP-to-Mel VC augmentation method is the same as the Mel-to-Mel VC system, and the loss function is the same as [22].

For the in-set speaker augmentation scenario, the word error rate (WER) and Cosine similarity are adopted as objective metrics to measure the VC and TTS systems. Fig. 4 and Table.3 shows the quality of synthesized speech from

Table 3. The WER[%] and cosine similarity for different system on the in-set experiment.

Model	Cosine/Utt Num. (average/all)	Utt Num. (> 0.6)	WER[%]
PPP-to-Mel(9utt)	0.555/68000	26029	9.11
Mel-to-Mel(9utt)	0.510/68000	23100	10.28
TTS (DiDi)	0.475/10000	4263(> 0.5)	-

different VC systems on in-set speakers' data. Each VC system generates 68000 text-dependent utterances (340*200=68000, each speaker generate 200 utterances). Comparing with the Mel-to-Mel system, the PPP-to-Mel system's average speaker similarity is higher. Moreover, as shown in Table.3, the WER of the PPP-to-Mel VC system is less than Mel-to-Mel in retained utterances data. Therefore, the speech quality of the PPP-to-Mel system is higher in terms of these objective metrics.

3.3 Data augmentation based on the TTS system

We also train a one-hot multi-speaker TTS system to generate the augmented data. The system is based on Tacotron-2 [22] with GMMv2 [1] attention. For the multi-speaker modeling, a naive embedding-table based strategy is employed, where 128 dimensional embeddings learned through model optimization are concatenated to the encoder's output sequence, guiding the attention mechanism and the decoder with target speaker's information.

The model is trained from the DiDi-speech [8] database with 500 speakers. For each pair of target speaker and desired keyword, we synthesize 20 speech samples with identical voice and lexical content.

3.4 Speaker augmentation based on speed perturbation

We use speed perturbation based on the SoX *speed* function that modifies the pitch and tempo of speech by resampling. This strategy has been successfully used in the speech and speaker recognition tasks [25, 13]. The limited text-dependent dataset is expanded by creating data created two versions of the original signal with speed factors of 0.9 and 1.1. The new classifier labels are generated at the same time since speech samples after pitch shift are considered to be from new speakers.

4 Experimental results

Table.2 shows the results of different data augmentation strategies. The evaluation metrics are Equal Error Rate (EER) and minimum Detection Cost Function (mDCF) with $P_{\text{target}} = 0.1$. The baseline system employs the original limited text-dependent dataset (9 utts per speaker) to fine tune the pre-trained model.

Since the size of in-set speaker dataset is too small, the system performance is degraded significantly. On the other hand, since the pitch shift AUG expand the number of speakers, the EER of the system has been improved by 10% relatively. The VC AUG with the PPP-to-Mel system also reduces the EER by relatively 20%. Moreover, it is observed that the system with both pitch shift AUG and VC AUG achieves the best performance. Experimental results show that, in the scenario of limited training data, the proposed method significantly reduces the EER from 6.51% to 4.48%, and the performance of the $\text{mDCF}_{0.1}$ also improves from 0.265 to 0.212.

Without the Pitch shift Aug, the VC AUG_{in} (PPP-to-Mel) have the lower EER and $\text{mDCF}_{0.1}$ than TTS Aug under the less speakers. Since all the synthesized speech sentences have the similar tone regarding the TTS Aug system, the VC Aug is more suitable than TTS Aug in the text-dependent speaker verification task.

Furthermore, since the speech quality and speaker similarity of synthesized speech from the PPP-to-Mel VC system are better than the Mel-to-Mel VC system, a better result is achieved by using the PPP-to-Mel VC system for data augmentation. Nevertheless, the Mel-to-Mel VC system explores the direction of out-of-set unseen speaker augmentation and achieves some improvement. The results obtained show that the VC Aug_{in} method is feasible, while the VC Aug_{out} method still needs to be explored in the future.

5 Conclusion

This paper proposes two voice conversion based data augmentation methods to improve the performance of text-dependent speaker verification systems under the limited training data scenario. The results show that VC-AUG and pitch-shift strategy are feasible and effective. In the future works, we will further explore the methods and strategies for voice conversion based data augmentation with unseen or even artificiality created speakers.

6 Acknowledgement

This research is funded in part by the Synear and Wang-Cai donation lab at Duke Kunshan University. Many thanks for the computational resource provided by the Advanced Computing East China Sub-Center.

References

1. Battenberg, E., Skerry-Ryan, R., Mariooryad, S., Stanton, D., Kao, D., Shannon, M., Bagby, T.: Location-relative attention mechanisms for robust long-form speech synthesis. In: Proc. ICASSP. pp. 6194–6198 (2020)
2. Cai, W., Chen, J., Li, M.: Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System. In: Proc. Odyssey. pp. 74–81 (2018)

3. Cai, Z., Zhang, C., Li, M.: From Speaker Verification to Multispeaker Speech Synthesis, Deep Transfer with Feedback Constraint. In: Proc. Interspeech (2020)
4. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep Speaker Recognition. In: Proc. Interspeech (2018)
5. Das, R.K., Kinnunen, T., Huang, W.C., Ling, Z.H., Yamagishi, J., Yi, Z., Tian, X., Toda, T.: Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions. In: Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020. pp. 99–120. https://doi.org/10.21437/VCC_BC.2020-15
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: Proc. CVPR. pp. 4685–4694 (2019). https://doi.org/10.1109/CVPR.2019.00482
7. Du, J., Na, X., Liu, X., Bu, H.: AISHELL-2: Transforming Mandarin ASR Research Into Industrial Scale. In: arXiv:1808.10583 (2018)
8. Guo, T., Wen, C., Jiang, D., Luo, N., Zhang, R., Zhao, S., Li, W., Gong, C., Zou, W., Han, K., Li, X.: Didispeech: A large scale mandarin speech corpus. arXiv:2010.09275 (2020)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proc.CVPR. pp. 770–778 (2016)
10. Heigold, G., Moreno, I., Bengio, S., Shazeer, N.: End-to-end text-dependent speaker verification. In: Proc. ICASSP. pp. 5115–5119 (2016). https://doi.org/10.1109/ICASSP.2016.7472652
11. Huadi Zheng, Cai, W., Tianyan Zhou, Shilei Zhang, Li, M.: Text-independent voice conversion using deep neural network based phonetic level features. In: Proc. ICPR. pp. 2872–2877 (2016). https://doi.org/10.1109/ICPR.2016.7900072
12. Huang, Y., Chen, Y., Pelecanos, J., Wang, Q.: Synth2aug: Cross-domain speaker recognition with tts synthesized speech. In: 2021 IEEE Spoken Language Technology Workshop (SLT). pp. 316–322 (2021). https://doi.org/10.1109/SLT48900.2021.9383525
13. Ko, T., Peddinti, V., Povey, D., Khudanpur, S.: Audio Augmentation for Speech Recognition. In: Proc. Interspeech. pp. 3586–3589. (2015)
14. Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W.Z., Sotelo, J., de Brébisson, A., Bengio, Y., Courville, A.C.: MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. In: Advances in Neural Information Processing Systems
15. Larcher, A., Lee, K.A., Ma, B., Li, H.: Text-Dependent Speaker Verification: Classifiers, Databases and RSR2015. Speech Communication **60** (05 2014). https://doi.org/10.1016/j.specom.2014.03.001
16. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: A Large-Scale Speaker Identification Dataset. In: Proc. Interspeech. pp. 2616–2620 (2017)
17. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: An ASR corpus based on public domain audio books. In: Proc. ICASSP. pp. 5206–5210 (2015). https://doi.org/10.1109/ICASSP.2015.7178964
18. Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., Khudanpur, S.: Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. In: Proc. Interspeech. pp. 3743–3747 (2018). https://doi.org/10.21437/Interspeech.2018-1417
19. Qin, X., Bu, H., Li, M.: HI-MIA: A Far-Field Text-Dependent Speaker Verification Database and the Baselines. In: Proc. ICASSP. pp. 7609–7613 (2020)

20. Qin, X., Cai, D., Li, M.: Far-Field End-to-End Text-Dependent Speaker Verification based on Mixed Training Data with Transfer Learning and Enrollment Data Augmentation. In: Proc. Interspeech (2019)
21. Qin, X., Li, M., Bu, H., Rao, W., Das, R.K., Narayanan, S., Li, H.: The INTERSPEECH 2020 Far-Field Speaker Verification Challenge. In: Proc. Interspeech. pp. 3456–3460 (2020). https://doi.org/10.21437/Interspeech.2020-1249
22. Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R.A., Agiomvrgiannakis, Y., Wu, Y.: Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In: Proc. ICASSP. pp. 4779–4783 (2018)
23. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: x-vectors: Robust DNN Embeddings for Speaker Recognition. In: Proc. ICASSP. pp. 5329–5333 (2018)
24. Wan, L., Wang, Q., Papir, A., Moreno, I.L.: Generalized End-to-End Loss for Speaker Verification. In: Proc. ICASSP. pp. 4879–4883 (2018). https://doi.org/10.1109/ICASSP.2018.8462665
25. Yamamoto, H., Lee, K.A., Okabe, K., Koshinaka, T.: Speaker Augmentation and Bandwidth Extension for Deep Speaker Embedding. In: Proc. Interspeech. pp. 406–410 (2019)
26. Yang, Y., Li, M.: The Sequence-to-Sequence System for the Voice Conversion Challenge 2020. In: Proc. Interspeech 2020 satellite workshop on Spoken Language Interaction for Mobile Transportation System
27. Yi, Z., Huang, W.C., Tian, X., Yamagishi, J., Das, R.K., Kinnunen, T., Ling, Z.H., Toda, T.: Voice Conversion Challenge 2020 âĂŞ- Intra-lingual semi-parallel and cross-lingual voice conversion âĂŞ-. In: Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020. pp. 80–98. https://doi.org/10.21437/VCC_BC.2020-14
28. Zhao, G., Ding, S., Gutierrez-Osuna, R.: Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams. In: Proc. Interspeech. pp. 2843–2847 (2019). https://doi.org/10.21437/Interspeech.2019-1778
29. Zhou, X., Ling, Z.H., King, S.: The Blizzard Challenge 2020. In: Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020. pp. 1–18 (2020). https://doi.org/10.21437/VCC_BC.2020-1