INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Improving robustness of one-shot voice conversion with deep discriminative speaker encoder

*Hongqiang Du, Lei Xie**

Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, China

hqdu@nwpu-aslp.org, lxie@nwpu.edu.cn

Abstract

One-shot voice conversion has received significant attention since only one utterance from source speaker and target speaker respectively is required. Moreover, source speaker and target speaker do not need to be seen during training. However, available one-shot voice conversion approaches are not stable for unseen speakers as the speaker embedding extracted from one utterance of an unseen speaker is not reliable. In this paper, we propose a deep discriminative speaker encoder to extract speaker embedding from one utterance more effectively. Specifically, the speaker encoder first integrates residual network and squeeze-and-excitation network to extract discriminative speaker information in frame level by modeling frame-wise and channel-wise interdependence in features. Then attention mechanism is introduced to further emphasize speaker related information via assigning different weights to frame level speaker information. Finally a statistic pooling layer is used to aggregate weighted frame level speaker information to form utterance level speaker embedding. The experimental results demonstrate that our proposed speaker encoder can improve the robustness of one-shot voice conversion for unseen speakers and outperforms baseline systems in terms of speech quality and speaker similarity.

Index Terms: voice conversion, one-shot, speaker embedding

1. Introduction

Voice conversion (VC) is a technique to modify the speech signal of a source speaker to make it sound like that of a target speaker without changing the linguistic content [1]. This technique has many applications, including expressive speech synthesis, speech enhancement, movie dubbing as well as other entertainment applications.

Various approaches have been proposed to achieve voice conversion, such as Gaussian mixture model (GMM) [2, 3, 4], frequency warping approaches [5, 6, 7], exemplar based methods [8, 9, 10], and neural network based methods [11, 12, 13, 14, 15]. While these works require to know either source speaker or target speaker or both in training, which limits their use in the real application scenarios. Recently, one-shot voice conversion approaches [16, 17, 18, 19] are proposed. Compared with previous methods, source and target speakers at run-time are not required to be seen during training. Additionally, only one utterance from the source speaker and target speaker respectively is needed. The speaker identity of converted speech can be controlled independently by the speaker embedding extracted from target speech.

Despite recent progress, the available one-shot voice conversion approaches are not stable for unseen speakers [20]. This is mainly because speaker embedding extracted from one utterance of an unseen speaker is not reliable [21, 20], which has a great influence on the stability of one-shot conversion [20]. Speaker embedding extractor can be a speaker encoder which is jointly trained with a conversion model or a pre-trained model for speaker information extraction, such as i-vector [22], d-vector [23], or x-vector [24]. The speaker embedding extractor jointly trained with the conversion model is more suitable for voice conversion than the pre-trained models [20]. When the network is jointly optimized, speaker embedding extractor is an inherent part of the model, which makes the generation of speech with correct speaker embedding consistently.

There are some studies on jointly training speaker encoder and voice conversion model. The speaker encoder generally consists of two parts: extracting frame level and utterance level features [25]. The frame level extractor takes acoustic features as input and outputs frame level features. It can be done via recurrent neural networks [26] and convolutional neural networks [18, 17]. In particular, convolutional neural network based residual network (ResNet) is a powerful speaker embedding extractor [27, 18]. The utterance level extractor further aggregates variable-length frame level features into utterance level speaker embedding. Average pooling [26, 18] is a popular method to obtain speaker embedding by averaging all frame level features. Another method [17] first uses the last state of unidirectional gated recurrent unit (GRU) layer as the utterance level speaker representation and then multi-head attention is utilized as a post-processing module to obtain final speaker embedding.

In this paper, to further improve the effectiveness of speaker embedding extracted from only one utterance of an unseen speaker, we propose a deep discriminative speaker encoder. Inspired by [28], first residual network and squeeze-and-excitation network [29] are integrated to extract discriminative frame level speaker information by modeling frame-wise and channel-wise interdependence in features. Then attention mechanism is introduced to give different weights to frame level speaker information. Finally, a statistic pooling layer [30] is used to aggregate weighted frame level speaker information to generate utterance level speaker embedding. Experimental results show that our proposed speaker encoder can improve the robustness of one-shot voice conversion and outperforms baseline systems in terms of speech quality and speaker similarity.

2. Related work

2.1. Residual network based speaker embedding extractor

Residual network (ResNet) has been widely used in speaker verification, which achieves promising performance for both long-duration and short-duration utterances [21, 27].

* corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-2132

The architecture of ResNet based speaker embedding extractor includes several ResNet blocks, followed by a statistics pooling layer and fully-connected (FC) layers. ResNet block operates on frame level features, which consists of convolutional layers, rectified linear units (ReLU) and batch normalization (BN) layers. Residual connection in ResNet block helps to build a deep neural network and avoids the vanishing gradient problem [31, 32]. Increasing the depth of a neural network can significantly improve the quality of representations [29]. Additionally, batch normalization helps to improve the stability of the training process of deep neural networks. Then a statistics pooling layer calculates the mean and standard deviation of each sample along the time-axis to form utterance level representation. Finally, two fully-connected (FC) layers project the utterance level representation into a fixed dimensional speaker embedding.

2.2. Squeeze-and-excitation network

Squeeze-and-excitation (SE) network [29] is first introduced to model channel interdependence in features for image classification. SE network can be used as a block and inserted in the convolutional neural network.

SE block consists of two operations: squeeze operation and excitation operation. Squeeze operation utilizes average pooling to generate channel-wise statistics. The statistics are mean vector z of frame level features h_t across the time-axis.

$$z = \frac{1}{T}\sum_{t}^{T} h_t \tag{1}$$

Then z is used in the excitation operation to calculate weights for each channel. The excitation operation is formulated as follows:

$$s = \sigma(W_2\delta(W_1 z)) \tag{2}$$

where $\sigma(\cdot)$ and $\delta(\cdot)$ are sigmoid and ReLU function respectively. $W_1 \in R^{C\times\frac{C}{r}}$, $W_2 \in R^{\frac{C}{r}\times C}$, C and r refer to the number of input channel and reduction ratio respectively. The channel-wise vector s contains the weight for each channel, which is between zero and one.

The final output $\hat{h}$ of SE block is obtained by channel-wise multiplication between the original input h and corresponding weight in s.

$$\hat{h} = sh \tag{3}$$

3. Robust one-shot voice conversion

3.1. Deep discriminative speaker encoder

Speaker encoder is an important component in the framework of one-shot voice conversion, which is directly related to performance of the whole network for unseen speakers [20].

Inspired by the previous study on ResNet and SE block, we integrate ResNet with SE block to build a deep discriminative speaker encoder (DDSE) for robust one-shot voice conversion. Figure 1 (a) depicts the framework of speaker encoder. It consists of frame level feature processing and utterance level feature processing.

The frame level feature processing part consists of a Conv2D layer, a ReLU function and a batch normalization (BN) layer, followed by four ResNet-SE blocks. The framework of ResNet-SE block is shown in Figure 1 (b). A ResNet-SE block mainly consists of convolution layers. Filters in the convolution layer explicitly model local features and allow spatial translation invariance, which make convolution layer suitable to extract frame level features [28]. SE block expands the temporal context of the frame level information by modeling channel interdependence in features, which has been verified to be helpful in speaker verification task [28]. The framework of SE block is shown in Figure 1 (c). An average pooling layer is utilized to generate channel-wise statistics. Then, two fully-connected layers capture the local channel dependencies. The first fully-connected layer can be used to reduce the feature dimension for controlling the computational cost, while the second fully-connected layer restores the number of feature to the original dimension. Finally, the channel-wise vector is obtained with a sigmoid layer to pay more attention to the discriminative channels for speaker representation.

The utterance level feature processing part consists of an attention block, followed by a statistic pooling layer and two fully-connected layers. Instead of directly using an average pooling layer where each frame level speaker information contributes equally to speaker embedding [18], we first introduce an attention block to further emphasize speaker related information. As shown in Figure 1 (d), the attention block takes the frame level speaker information as input and outputs the corresponding weights, which allows the speaker encoder to select the frames it deems relevant. Then a statistics pooling layer [25] is used to calculate the weighted means and weighted standard deviations of the final extracted frame level features. The mean and deviation are combined to form an utterance-level speaker representation. Finally, two fully-connected layers are introduced. The first one acts as a bottleneck layer to generate the low-dimensional speaker representation. The second one projects the speaker representation into a fixed dimensional speaker embedding.

In summary, convolution based ResNet is a powerful architecture to extract speaker representations by modeling relationships between frames. SE block contributes to the discriminative speaker representation learning by exploring the channel-wise information. Attention mechanism further makes speaker encoder emphasize speaker related information and overshadow other information. Therefore, the speaker embedding learned by this architecture concentrates on speaker characteristics more effectively.

3.2. Robust one-shot voice conversion with DDSE

The frameworks of available one-shot voice conversion approaches [18, 19] generally consist of a speaker encoder, a content encoder, and a decoder. AdaIN-VC [18] is a successful end-to-end implementation, which is relatively more robust for unseen speakers [20]. We use AdaIN-VC as a case study and replace the original speaker encoder with our proposed deep discriminative speaker encoder to make it more robust for unseen speakers. Note that the whole network is jointly optimized and the speaker encoder is optimized without explicit loss function.

The robust one-shot voice conversion method consists of two steps: conversion model training and run-time conversion. During the training stage, the speaker encoder and content encoder learn to extract speaker embedding and linguistic content representation from spectrum respectively. The decoder takes the two representations as inputs to reconstruct the spectrum. During run-time shown in Figure 2, the speaker encoder extracts utterance level speaker embedding from target speech. The content encoder extracts frame level content representation from

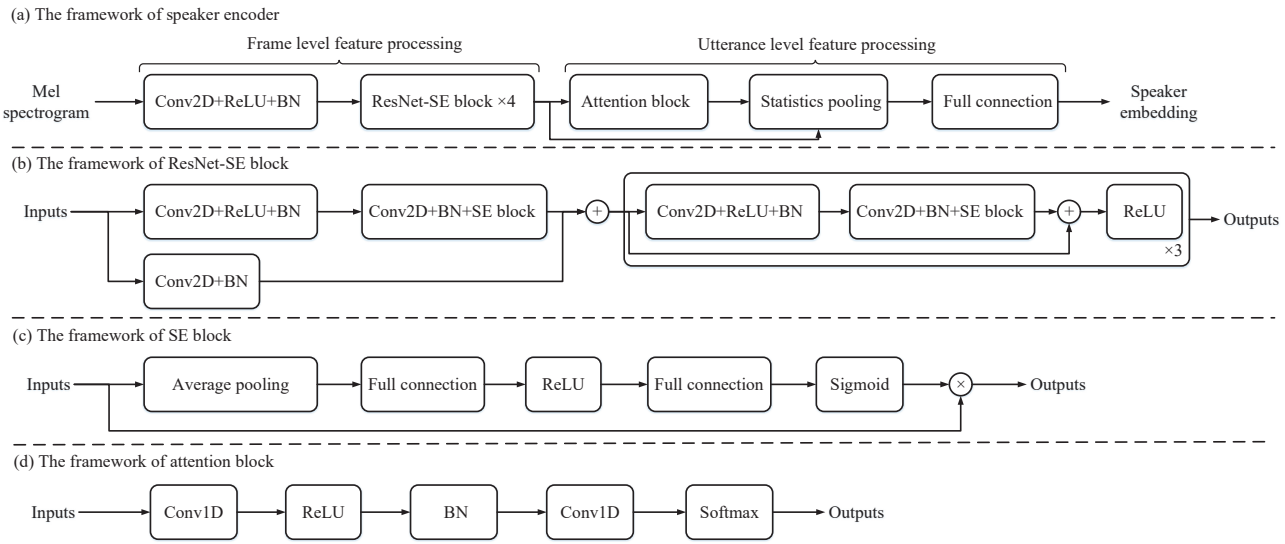

Figure 1: *The framework of the proposed deep discriminative speaker encoder for robust one-shot voice conversion. The speaker encoder consists of frame level feature processing and utterance level feature processing.*

source speech. The decoder takes content and speaker representations as input to reconstruct converted speech.

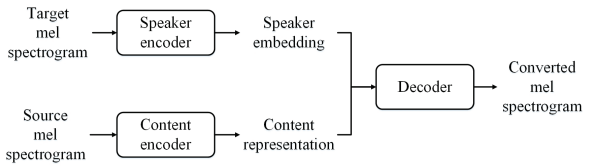

Figure 2: *The diagram of robust one-shot voice conversion at run-time conversion.*

4. Experimental setup

4.1. Database and feature extraction

CSTR-VCTK [33] database, containing 44 hours of speech from 109 speakers, is used to train the conversion model. Voice conversion experiments are carried out on CMU-ARCTIC [34] database and VCC 2016 dataset [35] respectively. For CMU-ARCTIC database, we select clb (female) and rms (male) as source speakers, and slt (female) and bdl (male) as target speakers. For VCC 2016 dataset, SF1 (female) and SM1 (male) are selected as source speakers, and TF2 (female) and TM3 (male) are selected as target speakers. For each target speaker, 20 utterances are used for evaluation. All audio files are downsampled to 16 kHz.

Librosa is employed to extract 256 dimensional mel spectrogram with 50ms frame length and 12.5ms frame shift.

4.2. Systems and setup

- VAE-ORI: This is the original AdaIN-VC [18] system. The speaker encoder and content encoder take 256 dimensional mel spectrogram as input and the output is 128 dimensional speaker embedding and content representation respectively. To further improve the speech quality, the auto-regressive technique is applied to the decoder.
- VAE-GSE: This has the same setting as VAE-ORI except that the speaker encoder is replaced with global speaker embedding (GSE) utilized in [17].
- VAE-ResNet: This has the same setting as VAE-ORI except that the speaker encoder is replaced with ResNet.
- VAE-DDSE: This has the same setting as VAE-ORI except that the speaker encoder is replaced with our proposed deep discriminative speaker encoder (DDSE). For the first ResNet-SE block, the kernel sizes and strides for the Conv2D layers are $\{3, 3, 1, 3, 3\}$ and $\{\{1, 1\}, \{1, 1\}, \{1, 1\}, \{1, 1\}, \{1, 1\}\}$ respectively. For the remaining ResNet-SE blocks, the kernel sizes and strides for the Conv2D layers are $\{3, 3, 1, 3, 3\}$ and $\{\{2, 2\}, \{1, 1\}, \{2, 2\}, \{1, 1\}, \{1, 1\}\}$ respectively. For the SE block, the reduction ratio r is set to 8.

Parallel WaveGAN [36] is used to synthesize the converted speech. We follow the original configurations.

5. Evaluations

5.1. Objective evaluation

Mel-cepstral distortion (MCD) [4] is employed to measure the spectral distortion. MCD is the Euclidean distance of the mel spectrogram between the converted speech and the target speech. Given a speech frame, the MCD is defined as follows:

$$\text{MCD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{n=1}^{N}\left(X_n^{conv} - X_n^{ref}\right)^2}, \quad (4)$$

where X_n^{conv} and X_n^{ref} are the n^{th} coefficient of the converted and target mel spectrogram, N is the dimension of mel spectrogram. The lower MCD indicates the smaller distortion.

Table 1 shows the average MCD for different systems. Intra-gender conversion is better than inter-gender conversion

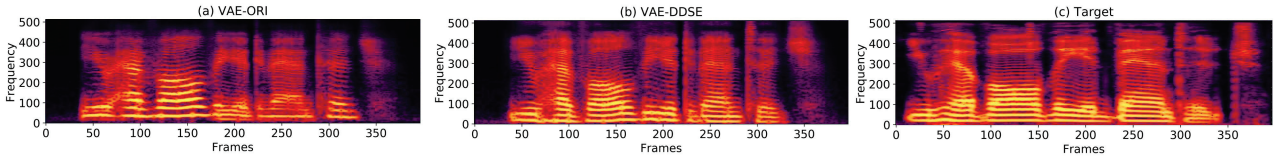

Figure 3: *Example of spectrum of the same utterance converted from clb (female) to bdl (male) by different systems: (a) VAE-ORI, (b) VAE-DDSE, and (c) Target.*

for the four systems. We also observe that VAE-DDSE significantly outperforms the VAE-ORI and VAE-GSE in both intra- and inter-gender conversions. VAE-DDSE performs better than VAE-ResNet and achieves the lowest average MCD of 10.75 dB. The objective evaluation results further confirm that the extracted speaker embedding has a great impact on the performance of one-shot voice conversion.

Table 1: *Comparison of average MCD (dB) for different systems.*

System	Inter	Intra	Average
VAE-ORI	13.02	12.85	12.93
VAE-GSE	15.66	15.52	15.59
VAE-ResNet	10.95	10.81	10.88
VAE-DDSE	10.81	10.70	10.75

In Figure 3, we show an example that compares spectrum of the same utterance converted from clb (female) to bdl (male) by different systems: (a) VAE-ORI, (b) VAE-DDSE, and (c) Target. The harmonics of the spectrum are closely related to the speaker identity [37], which is controlled by the extracted speaker embedding. The harmonics in Figure 3 (a) are clearly higher than that in Figure 3 (c), which indicates that the converted speech is not stable and speaker similarity is degraded. The harmonics maintain roughly the same between Figure 3 (b) and Figure 3 (c).

5.2. Subjective evaluation

For subjective evaluation, we first conduct AB and XAB preference tests to assess speech quality and speaker similarity. Then the mean opinion score (MOS) is utilized to evaluate speech naturalness. Each listener is asked to give an opinion score on a five-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). For each system, 20 samples are randomly selected from the 160 converted samples for listening tests. Ten listeners participated in all listening tests. Different listeners may listen to different samples. Listening tests cover all the 160 evaluation samples.

The subjective results of AB tests are presented in Figure 4. It is observed that our proposed VAE-DDSE significantly outperforms VAE-GSE and VAE-ORI.

Figure 5 shows the similarity preference results of XAB tests. As shown in Figure 5 (a) and (b), we observe that VAE-DDSE outperforms VAE-GSE and VAE-ORI respectively in terms of speaker similarity.

Figure 6 shows the mean opinion scores for different systems. Benefiting from the discriminative speaker embedding extracted from DDSE, VAE-DDSE achieves the highest MOS score. The synthesized samples can be found on the website [1].

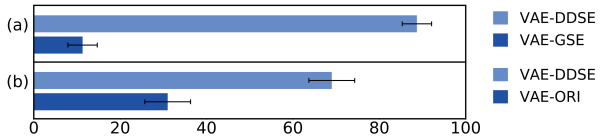

Figure 4: *Quality preference tests of converted speech samples with 95% confidence intervals for different systems.*

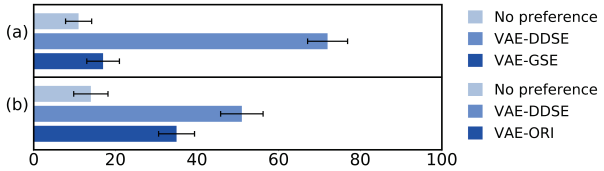

Figure 5: *Similarity preference tests of converted speech samples with 95% confidence intervals for different systems.*

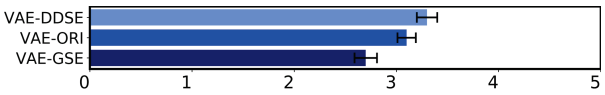

Figure 6: *Comparison of mean opinion scores with 95% confidence intervals for different systems*

6. Conclusions

In this study, we propose a deep discriminative speaker encoder to improve the robustness of one-shot voice conversion for unseen speakers. The speaker encoder first integrates residual network and squeeze-and-excitation network to extract frame level speaker information from time-axis and channel-axis. Then attention mechanism is used to further focus on the speaker related information. Finally a statistic pooling layer is used to aggregate weighted frame level speaker information to form utterance level speaker embedding. The experimental results demonstrate that our proposed speaker encoder can improve the robustness of one-shot voice conversion for unseen speakers in terms of speech quality and speaker similarity.

7. References

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[2] H. Benisty and D. Malah, "Voice conversion using GMM with enhanced global variance," in *Annual Conference of the International Speech Communication Association*, 2011.

[1] `https://dhqadg.github.io/robust/`

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on speech and audio processing*, vol. 6, no. 2, pp. 131–142, 1998.

[4] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[5] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2009.

[6] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2011.

[7] X. Tian, Z. Wu, S. W. Lee, N. Q. Hy, E. S. Chng, and M. Dong, "Sparse representation for frequency warping based voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 4235–4239.

[8] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *IEEE Spoken Language Technology Workshop*. IEEE, 2012, pp. 313–317.

[9] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[10] X. Tian, S. W. Lee, Z. Wu, E. S. Chng, and H. Li, "An exemplar-based approach to frequency warping for voice conversion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1863–1876, 2017.

[11] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 4869–4873.

[12] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2016, pp. 1–6.

[13] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[14] Z. Wang, W. Ge, X. Wang, S. Yang, W. Gan, H. Chen, H. Li, L. Xie, and X. Li, "Accent and speaker disentanglement in many-to-many voice conversion," in *International Symposium on Chinese Spoken Language Processing*. IEEE, 2021, pp. 1–5.

[15] X. Tian, Z. Wang, S. Yang, X. Zhou, H. Du, Y. Zhou, M. Zhang, K. Zhou, B. Sisman, L. Xie, and H. Li, "The NUS & NWPU system for voice conversion challenge 2020," in *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 2020, pp. 170–174.

[16] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, and H. Meng, "Voice conversion across arbitrary speakers based on a single target-speaker utterance." in *Annual Conference of the International Speech Communication Association*, 2018, pp. 496–500.

[17] H. Lu, Z. Wu, D. Dai, R. Li, S. Kang, J. Jia, and H. Meng, "One-shot voice conversion with global speaker embeddings." in *Annual Conference of the International Speech Communication Association*, 2019, pp. 669–673.

[18] J.-c. Chou and H.-Y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *Annual Conference of the International Speech Communication Association*, pp. 664–668, 2019.

[19] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*, 2019, pp. 5210–5219.

[20] T.-h. Huang, J.-h. Lin, and H.-y. Lee, "How far are we from robust voice conversion: A survey," in *IEEE Spoken Language Technology Workshop*. IEEE, 2021, pp. 514–521.

[21] N. Li, D. Tuo, D. Su, Z. Li, D. Yu, and A. Tencent, "Deep discriminative embeddings for duration robust speaker verification." in *Annual Conference of the International Speech Communication Association*, 2018, pp. 2262–2266.

[22] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[23] R. Doddipatla, N. Braunschweiler, and R. Maia, "Speaker adaptation in DNN-based speech synthesis using d-vectors." in *Annual Conference of the International Speech Communication Association*, 2017, pp. 3404–3408.

[24] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 5329–5333.

[25] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *Annual Conference of the International Speech Communication Association*, pp. 2252–2256, 2018.

[26] Y. Zhou, X. Tian, R. K. Das, and H. Li, "Many-to-many cross-lingual voice conversion with a jointly trained speaker embedding network," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2019, pp. 1282–1287.

[27] A. Gusev, V. Volokhov, T. Andzhukaev, S. Novoselov, G. Lavrentyeva, M. Volkova, A. Gazizullina, A. Shulipa, A. Gorlanov, A. Avdeeva, A. Ivanov, A. Kozlov, T. Pekhovsky, and Y. Matveev, "Deep speaker embeddings for far-field speaker recognition on short utterances," in *The Speaker and Language Recognition Workshop*, 2020, pp. 179–186.

[28] J. Xu, X. Wang, B. Feng, and W. Liu, "Deep multi-metric learning for text-independent speaker verification," *Neurocomputing*, vol. 410, pp. 394–400, 2020.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[30] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Annual Conference of the International Speech Communication Association*, 2017, pp. 999–1003.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[32] D. Garcia-Romero, A. McCree, D. Snyder, and G. Sell, "Jhu-hltcoe system for the voxsrc speaker recognition challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 7559–7563.

[33] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research*, 2017.

[34] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *ISCA workshop on speech synthesis*, 2004.

[35] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," *Annual Conference of the International Speech Communication Association*, pp. 1632–1636, 2016.

[36] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 6199–6203.

[37] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *arXiv preprint arXiv:1704.04222*, 2017.