DGC-VECTOR: A NEW SPEAKER EMBEDDING FOR ZERO-SHOT VOICE CONVERSION

Ruitong Xiao[1][*][†], Haitong Zhang[2][*], Yue Lin[2]

[1]South China University of Technology, China;[2] Netease Games AI Lab, China
[1]eerxiao@mail.scut.edu.cn; [2]{zhanghaitong01, gzlinyue}@corp.netease.com

ABSTRACT

Recently, more and more zero-shot voice conversion algorithms have been proposed. As a fundamental part of zero-shot voice conversion, speaker embeddings are the key to improving the converted speech's speaker similarity. In this paper, we study the impact of speaker embeddings on zero-shot voice conversion performance. To better represent the characteristics of the target speaker and improve the speaker similarity in zero-shot voice conversion, we propose a novel speaker representation method in this paper. Our method combines the advantages of D-vector, global style token (GST) based speaker representation and auxiliary supervision. Objective and subjective evaluations show that the proposed method achieves a decent performance on zero-shot voice conversion and significantly improves speaker similarity over D-vector and GST-based speaker embedding.

Index Terms— voice conversion, zero-shot, speaker embedding, global style token

1. INTRODUCTION

Voice conversion (VC) task is to convert a source speaker's voice to a target speaker's voice without changing the linguistic content [1]. It is widely applied in many fields such as entertainment, creative industry, and spoofed speech generation.

In terms of learning methods, previous works of VC can be roughly divided into two categories: automatic speech recognition (ASR) based approach and auto-encoder based approach. The ASR-based VC first extracts the phoneme posteriorgram (PPG) from the utterance by a pre-trained ASR model as the content features and then performs voice conversion using PPG. PPG is considered a kind of speaker-agnostic feature. ASR-based VC methods, such as [2], are widely used, but the performances highly depend on the performances of ASR models. Different from ASR-based VC, auto-encoder-based VC methods, such as [3], extract the content features through the content encoder and then perform voice conversion on the content features. Auto-encoder-based VC systems need to disentangle the speaker identity from the content information. Besides that, StarGAN-VC [4] and CycleGAN-VC [5] are two typical methods. Most of the previous VC works can perform VC on the seen speaker data, namely seen-to-seen VC, but are infeasible to few-shot VC. Thus, their scope of application is restricted. Few-shot VC is more challenging than seen-to-seen VC and has become a hot spot of current research. To overcome the restraint, some methods with a strong generalization ability have been proposed for the few-shot or even zero-shot VC task.

In the case of few-shot VC, speaker embedding is an important factor that can not be ignored since it can help VC models use limited training data more efficiently and generate voice more similar to the target speaker. Usually, the speaker embedding is generated by a pre-trained speaker encoder or a jointly-optimized speaker encoder. In Voice Conversion Challenge (VCC) 2020 [6], two kinds of speaker embedding are widely used. Some teams like [7] and [8] simply used the one-hot embedding as the speaker embedding. Automatic speaker verification (ASV) based speaker embedding was used in many schemes [9, 10].

Compared with few-shot VC, zero-shot VC algorithms do not need training the VC model with data from the target speaker. Therefore, the speaker embedding of zero-shot VC algorithms should have the ability to represent the characteristics of the target speaker. For example, VQVC [11] proposes to separate the speaker and content information through discrete codes and uses the residual feature for speaker representation. AUTOVC [12] uses D-vector [13, 14] as the speaker embedding. Chou et al. propose a zero-shot voice conversion system with instance normalization (AdaIN-VC) [15], which jointly optimizes a speaker encoder with the conversion model.

In this paper, we study the impact of speaker embeddings on zero-shot VC performance and aim to improve speaker similarity of zero-shot VC. In order to represent the characteristics of target speakers better and improve the speaker similarity, we propose a novel speaker representation method, namely DGC-vector. Our method combines the advantages of D-vector, global style token [16] based speaker representation and auxiliary supervision. A GST-based module with auxiliary speaker classifier is applied to generate speaker embedding from bottleneck feature of D-vector extractor. Objective and subjective evaluations show that our proposed method achieves a decent performance on zero-shot voice conversion and improves the speaker similarity significantly over other baseline speaker representations.

The rest of the paper is organized as follows. Section 2 simply introduces related works. Section 3 introduces the proposed method in detail. Section 4 gives the experimental results and the corresponding analysis. Finally, the paper is concluded.

2. RELATED WORK

2.1. D-vector

D-vector [13] is a typical kind of speaker embedding extracted from the deep neural network based ASV system. Due to its outstanding performance in speaker verification, D-vector has been widely used to build a speaker adaptive model for both VC and text-to-speech (TTS) [12, 17]. [14] proposes a generalized end-to-end loss for speaker verification, which achieved the SOTA results in speaker verification. The proposed system has been applied in some VC systems, such as AUTOVC [12].

Regarding speaker verification, D-vector has proved to be generalizable to the unseen speakers to some extent [14]. However,

---

speaker embedding extracted from speaker verification systems may not adapt to few-shot or zero-shot VC well, leading to a poor speaker similarity between the target voice and the converted ones. Fig. 2 shows the 2D t-SNE visualization of speaker embeddings computed from target speakers' ground truth utterances and converted utterances. The first left column corresponds to AUTOVC conditioned on D-vector extracted from [14]. We define converted utterances and their corresponding target speakers' utterances as a conversion group. As shown in the upper part of the first left column in Fig. 2, the converted utterances are close to their target speakers' utterances, and different conversion groups are well separated. It indicates that D-vector adapts to AUTOVC well in seen-to-seen VC. However, as shown in the bottom part, some converted utterances are far from target speakers' utterances in zero-shot VC. It illustrates that the speaker similarity significantly decreases in the zero-shot conversion, and the model does not adapt well to the unseen speakers.

2.2. Global Style Token

Global style token (GST) [16] is able to extract a sentence-level representation from the reference utterance. It consists of a reference encoder and a style token module. In the original paper [16], GST is used to capture the speaking style. In the M2VoC challenge [18], Wang et al. [19] used GST to represent speaker and prosody characteristics. For VC, Lu et al. [20] proposes to use the GST module as a speaker encoder. In this paper, we refer this kind of speaker embedding in [20] as G-vector. G-vector can adapt well the VC model since the GST module is jointly optimized with the whole VC model. However, G-vector is extracted without any explicit supervision. Moreover, the GST module is trained with only the VC training dataset instead of a large multi-speaker dataset typically used in training ASV systems. Thus, the generalization ability of the G-vector may be limited.

3. PROPOSED METHOD

In this paper, we use AUTOVC [12] as the framework. Because we focus on the impact of speaker embedding, we keep the content encoder and decoder settings unchanged. Under the framework of AUTOVC, we propose a novel speaker embedding extraction method, which combines the advantages of the ASV-based method and the GST-based method.

3.1. Framework

AUTOVC [12] is an auto-encoder VC model with a carefully designed bottleneck (See framework part of Fig. 1). AUTOVC is trained with the self-reconstruction loss like CVAE, but it has the distribution matching property similar to GAN. The loss function of AUTOVC is defined as

$$L_{rec} = \left\| x - x' \right\|_2 + \left\| x - x'' \right\|_2 + \left\| E_c(x) - E_c(x'') \right\|_1 \quad (1)$$

where E_c denotes content encoder. The fundamental idea of AUTOVC is that the correctly-designed bottleneck learns to disentangle the content information and speaker information by introducing carefully tuned dimension reduction and temporal downsampling. This simple scheme leads to a significant performance improvement. As a result, AUTOVC achieves superior performances on the many-to-many VC task and decent performances on the zero-shot VC task.
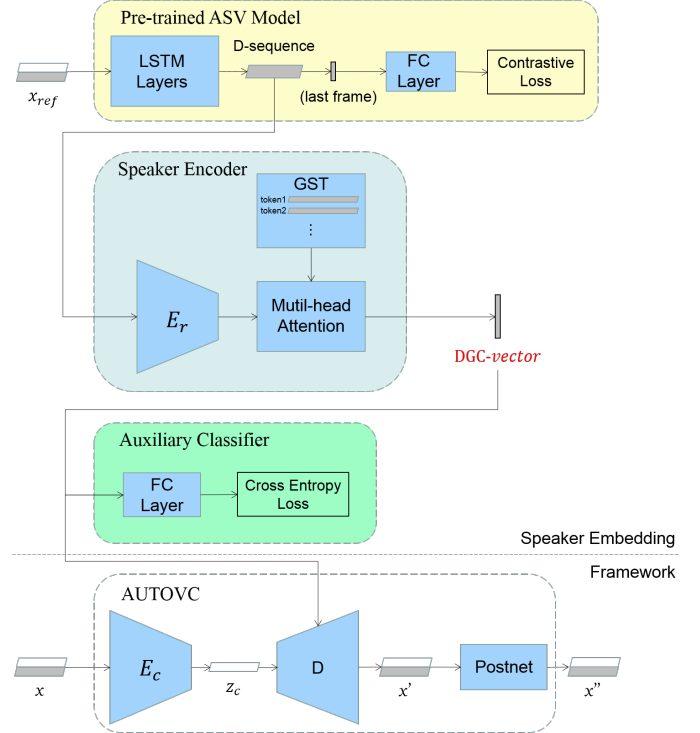

Fig. 1. The architecture of entire model. Proposed speaker embedding method is shown in the upper part. The fixed LSTM layers extracts D-sequence as input of speaker encoder and then speaker encoder extracts speaker embedding, namely DGC-vector.

3.2. DGC-vector

This paper proposes a novel speaker representation for zero-shot voice conversion, which combines the advantages of both the ASV-based method, the GST-based method and auxiliary classification supervision. We call it DGC-vector. Our proposed method consists of three parts: a pre-trained ASV model, a GST-based speaker encoder and an auxiliary classifier.

3.2.1. D-sequence

D-vector is a widely used kind of ASV method. Although D-vector can achieve good results in speaker recognition, directly using D-vector does not adapt well to zero-shot VC because the pre-trained ASV model and the conversion model are not trained jointly using the conversion optimization function. Inspired by the successful application of phonetic posteriorgrams in VC, we use a pre-trained ASV model to extract the bottleneck features, which we called D-sequence in this paper. The pre-trained ASV model is trained following setting of [14]. As shown in pre-trained ASV model of Fig. 1, the ASV model consists of several LSTM layers and a fully connected (FC) layer. We defined the output of the last LSTM layer as the D-sequence. D-sequence retains time dimension and we believe that D-sequence contains as much speaker information as D-vector and retain more information to adapt to the conversion model. The reason why we choose [14] is that it gets an outstanding performance in ASV task. Indeed, we can flexibly choose other effective method as pre-trained ASV model. The key is that we hope to introduce priori knowledge of ASV task into speaker embedding through pre-trained
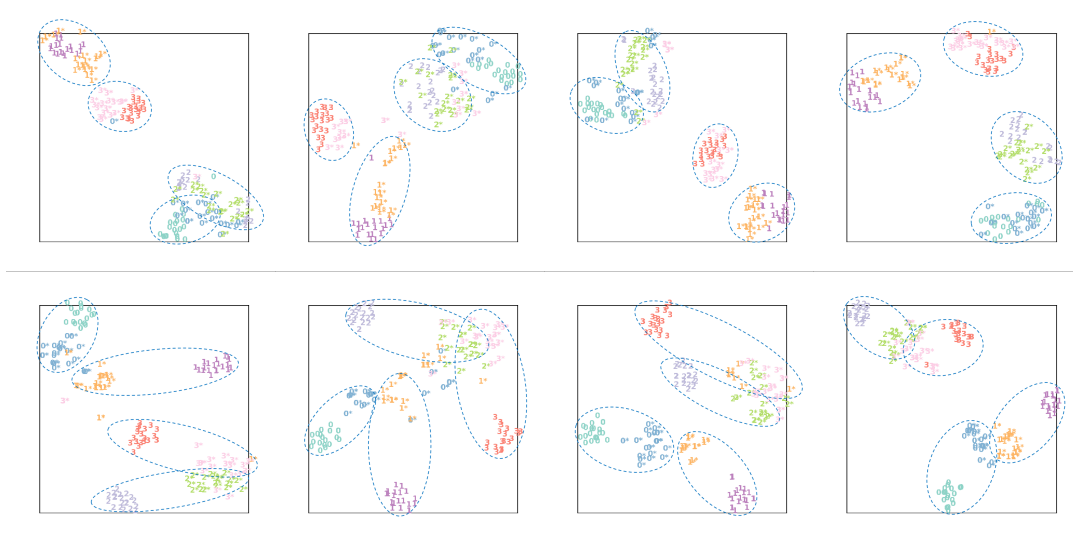
Fig. 2. The visualization of speaker similarity by t-SNE. The upper and lower rows correspond to seen-to-seen VC and zero-shot VC, respectively. From left to right, four columns correspond to D-vector, G-vector, DG-vector and DGC-vector, respectively. Each small number in the graph is an utterance sample. Different numbers without asterisk correspond to different target speakers; numbers with an asterisk refer to converted utterances corresponding to the target speaker whom the number stands for. We roughly mark each conversion pair with a blue dotted circle.

ASV model.

3.2.2. GST-based Speaker Encoder

In order to generate speaker embedding adapted to the VC task, we jointly optimize a GST-based speaker encoder with the conversion model. GST-based speaker encoder has been proved to be effective in extracting a global speaker embeddings for VC [20]. This part can also be replaced by other effective models, because what we need is a module which can jointly optimize with conversion model and extract speaker embedding from input feature. Instead of taking mel-spectrogram as the GST-based speaker encoder's input [16, 18, 20], we use the D-sequence described in section 3.2.1 as the input.

The speaker encoder includes two parts, namely a reference encoder E_r and a GST module, as shown in Fig. 1. We assume that the speaker information is time-invariant. Thus we use the reference encoder to extract a fixed-dimension vector. The reference encoder comprises three convolution layers and a GRU[21] layer, and we use the final output of the GRU layer as the input of the GST module. A set of randomly initialized learnable vectors is used as GSTs. A multi-head attention module takes the GRU's output as the query and GSTs as the key and value. Through matrix multiplication, we eventually get the final speaker embedding.

3.2.3. Auxiliary classification task

To ensure the speaker encoder extracts the speaker-related information, we add an auxiliary classification task as shown in Fig. 1. The classifier includes only one fully connected (FC) layer, which takes the speaker embedding from the GST module as input and predicts its speaker. The classification loss is defined as follows,

$$L_{class} = \mathbb{E}_{x_{ref}\epsilon\chi}\left[-log(C_a(c|E_s(x_{ref})))\right] \quad (2)$$

where C_a, E_s, χ and c represent auxiliary classifier, entire speaker embedding module, training dataset and speaker identity of input

x_{ref}, respectively. If we only use the self-reconstruction loss like [20], the GST-based speaker encoder may learn some information that is not related to speaker identity [16]. Thus it is likely to make the speaker encoder lose its desired function and be harmful for its generalization to the unseen speakers. Therefore, we use the classification task to enforce explicit supervision on the speaker encoder to improve its generalization.

Finally, the objective function of the whole model is defined as

$$L = \lambda_{rec}L_{rec} + \lambda_{class}L_{class} \quad (3)$$

where λ_{rec} and λ_{class} are the weight items for respective loss.

4. EXPERIMENT AND EVALUATION

4.1. Dataset and experiment setup

In the following experiments, we compared different AUTOVC variants with different speaker embeddings, including D-vector, G-vector, DG-vector (DGC-vector without classification supervision), and DGC-vector. All models were trained on data of 80 speakers randomly selected from CSTR VCTK corpus [22]. Each speaker has about 400 sentences. We used four target speakers and four source speakers in the seen-to-seen VC scenario. In the zero-shot VC scenario, we used the VCC2020 dataset [6]. We used TEF1, TEF2, TEM1, and TEM2 as target speakers and SEF1, SEF2, SEM1, and SEM2 as source speakers. 'T'/'S' is for target/source, 'E' is for English, 'F'/'M' is for female/male. In each VC task, we had 16 conversion pairs, including female to female (F2F), male to male (M2M), female to male (F2M), and male to female (M2F). Each model generated 25 converted samples for each pair.

80-dimension mel-spectrogram is extracted as the input acoustic feature, and the input frame length is 160. Thus the size of input x is 80×160. Speaker embedding generated from different methods is 256-dimension. λ_{rec} and λ_{class} are set to 1 and 0.5, respectively. Other model configuration settings are the same as that of AU-

Table 1. Result comparison for zero-shot VC among different speaker embedding methods. 'Avg', 'Inter', and 'Intra' mean the average, inter-gender conversion, and intra-gender conversion, respectively.

Methods	MCD			MAE of F0		
	Avg	*Inter*	*Intra*	*Avg*	*Inter*	*Intra*
D-vector	11.54	11.67	11.42	22.85	26.66	19.05
G-vector	11.54	11.7	11.39	22.97	28.31	**17.63**
DG-vector	**11.32**	**11.48**	**11.16**	22.27	26.04	18.49
DGC-vector	11.46	11.51	11.41	**20.5**	**23.03**	17.97

Fig. 3. MOS of similarity for naturalness (upper) and similarity (bottom). Error bars denote the standard deviation of samples.

TOVC [12]. A pre-trained Hifi-GAN [23] vocoder was used as the vocoder to reconstruct the waveform conditioned on the predicted mel-spectrogram by the conversion model.

4.2. Objective evaluation

To visualize speaker similarity of utterances, we plotted the D-vector of the converted utterances and target speakers' utterances in 2D space by t-SNE [24]. As Fig. 2 illustrates, the converted utterances generated by the model with DGC-vector are close to the respective target speakers' utterances in the seen-to-seen case, and different conversion pairs are well separated. In the zero-shot VC, distances between the converted and target utterances are shorter than in other methods. It indicates that the proposed method can improve the speaker similarity and generalization ability in the zero-shot VC task. In addition, the converted utterances gather in a more dense circle, which shows that the speaker consistency of the converted utterances generated by the proposed method is the best. However, the model with G-vector performed well neither in the seen-to-seen scenario nor the zero-shot scenario. As mentioned in section 2.1, the model with D-vector performed significantly worse in the zero-shot scenario. The model with DG-vector suffers from a similar problem.

Dynamic time warping Mel cepstral distortion (DTW-MCD) [25] and mean absolute error (MAE) of F0 were used as the objective metric, and the lower the better. We used the converted utterances and the ground truth data to calculate the metrics. Because we only got the ground truth data for the VCC2020 dataset, we only computed the results for zero-shot VC. The results are shown in Table 1. In terms of MCD, the proposed method gets a slight improvement over D-vector and G-vector. For MAE of F0, the proposed method is significantly lower than other methods, which indicates that the proposed method has a better F0 similarity which is related to voice characteristics of speakers.

4.3. Subjective evaluation

For the subjective evaluation, we used the mean opinion score (MOS) test in zero-shot VC scenario. We used 5 randomly selected utterances of 25 generated converted samples for each conversion pair, so each method had 80 (16×5) utterances in all. Thirteen Chinese raters who are proficient in English evaluated the naturalness and speaker similarity of utterances using scores ranging from 1 to 5; the higher the better. The demo can be found at https://shaw0fr.github.io/DGC-vector-DEMO/.

In the naturalness MOS test, we notice that different methods give comparable average performances as shown in the upper part of Fig. 3. This result is unsu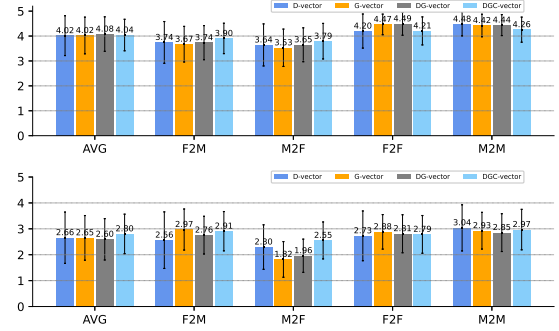rprising since we only made modifications to the speaker representation. However, we find that the naturalness MOS of D-vector, G-vector and DG-vector get worse during intergender VC. The problem may be related to generalization ability of methods. And DGC-vector is more stable. What is surprising is that the score are much higher than previous work [12]. We suspect that our raters are not native speakers and not aware of nuance in the converted speech.

In the speaker similarity MOS test, we evaluated speaker similarity in the zero-shot VC task. As the bottom part of Fig. 3 declares, our method got the highest average score. Compared with G-vector and DG-vector, we find that D-sequence does not bring expected benefit, although the performance of G-vector is more unstable for different conversion pairs and significantly gets worse for male-to-female VC. We suspect that the GST-based speaker encoder may not make full use of the speaker information contained by D-sequence. Compared with D-vector and DG-vector, DG-vector did not give much improvement, which also indicates the unsupervised GST speaker encoder does not leverage the extra helpful information in D-sequence. However, with the classification supervision, speaker similarity remained relatively high score for different conversion pairs, which indicates DGC-vector has the best generalization ability. Moreover, DGC-vector is more stable than other methods, similar to the situation in naturalness MOS test. Especially during inter-gender VC, whether it is male-to-female or female-to-male, our method can give a good performance. These conclusions are consistent with the conclusions obtained from Fig. 2 in section 4.2. Our experiments show that the improvement of DGC-vector effect is not simply the accumulation of positive benefits of multiple modules, because the improvement of DG-vector is not obvious. In order to make full use of the advantages of D-sequence and GST, auxiliary supervision is needed. Conclusively, although results of naturalness MOS are similar, our method can significantly improve the speaker similarity and generalization ability.

5. CONCLUSIONS

In this paper, we study the impact of speaker embedding on zero-shot VC effects. In order to represent the characteristics of target speaker better and improve speaker similarity of zero-shot VC, we propose a novel speaker embedding method. Under the framework of AU-TOVC, a GST-based module with auxiliary supervision is applied to generate speaker embedding from bottleneck feature of ASV model. Objective and subjective evaluations show that the proposed method can improve speaker similarity and generalization ability.

6. REFERENCES

[1] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou LI, "An overview of voice conversion and its challenge: from statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2021.

[2] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016, pp. 1–6.

[3] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[4] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion," in *INTERSPEECH*, 2019.

[5] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6820–6824.

[6] Zhao Yi, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhen-Hua Ling, and Tomoki Toda, "Voice Conversion Challenge 2020 — Intralingual semi-parallel and cross-lingual voice conversion —," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 80–98.

[7] Haitong Zhang, "The NeteaseGames System for Voice Conversion Challenge 2020 with Vector-quantization Variational Autoencoder and WaveNet," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 175–179.

[8] Oriol Barbany and Milos Cernak, "FastVC: Fast Voice Conversion with non-parallel data," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 145–149.

[9] Qiuyue Ma, Ruolan Liu, Xue Wen, Chunhui Lu, and Xiao Chen, "Submission from srcb for voice conversion challenge 2020," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 131–135.

[10] Li-Juan Liu, Yan-Nian Chen12, Jing-Xuan Zhang, Yuan Jiang12, Ya-Jun Hu, Zhen-Hua Ling, and Li-Rong Dai, "Non-parallel voice conversion with autoregressive conversion model and duration adjustment," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 126–130.

[11] Da-Yi Wu and Hung-Yi Lee, "One-shot voice conversion by vector quantization," in *45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7734–7738.

[12] 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.

[13] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[14] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[15] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," in *INTERSPEECH*, 2019.

[16] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[17] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[18] Qicong Xie, Xiaohai Tian, Guanghou Liu, Kun Song, Lei Xie, Zhiyong Wu, Hai Li, Song Shi, Haizhou Li, Fen Hong, et al., "The multi-speaker multi-style voice cloning challenge 2021," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8613–8617.

[19] Tao Wang, Ruibo Fu, Jiangyan Yi, Jianhua Tao, Zhengqi Wen, Chunyu Qiang, and Shiming Wang, "Prosody and voice factorization for few-shot speaker adaptation in the challenge m2voc 2021," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8603–8607.

[20] Hui Lu, Zhiyong Wu, Dongyang Dai, Runnan Li, Shiyin Kang, Jia Jia, and Helen Meng, "One-shot voice conversion with global speaker embeddings.," in *INTERSPEECH*, 2019, pp. 669–673.

[21] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *EMNLP*, 2014.

[22] Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[23] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[24] Laurens, Van Der Maaten, and Hinton, Geoffrey , "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 2605, pp. 2579–2605, 2008.

[25] Robert F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 1993, vol. 1, pp. 125–128.