Voice Style Transfer Based on Improved CycleGAN Network

Ling Lei*, Ruomu Wei, Xinran Wu

School of Information and Communication Engineering, Communication University of China, Beijing, 100024, China

*leiling@cuc.edu.cn

Keywords: CycleGAN; Voice style transfer; Missing frame filling; Second adversarial losses

Abstract: Voice style transfer refers to transferring the timbre style of the source speaker's voice to the tonal style of the target speaker while keeping the speech content intact. Deep learning technology has promoted voice technology's advancement and large-scale application in recent years. Among them, the CycleGan network, used for the first time in image transformation, also shows advantages in voice style transfer tasks. However, during the speech type conversion of the CycleGan network, the generated voice quality is often low, and the effect is not good, so based on this, this paper proposes three methods for improvement. In particular, a second adversarial loss is introduced to alleviate the problem of over-smoothing in statistical models. The generator and discriminator structures are optimized, and the inputs are optimized using 2D-1D-2D convolutional structures and PatchGAN, optimizing input feature details and reducing spectral distortion. In addition, auxiliary technology Missing Frame Fill (FIF), is applied to make the model pay more attention to the time-frequency structure of the sound. Then, based on the AISHELL-3 dataset, the traditional CycleGAN and the improved CycleGAN network were used to conduct tests on the voice style transfer, respectively. The test results show that compared with the traditional CycleGAN network, the improved CycleGAN network has achieved significant improvement in the subjective evaluation indicators of voice naturalness and similarity scores, as well as the objective indicators MCD and MSD, which verifies the effectiveness of the above three improvement measures.

1. Introduction

Deep learning technology has recently fueled voice technology's advancement and wide application. In digital voice processing, voice style transfer has gradually become an important research direction with different needs and scenarios, such as vocal processing for music creation, voice generation and conversion for intelligent voice assistants, voiceover processing for movies and TV series, and voice desensitization in secret environments.

Traditional speech-type transmission methods are mainly based on classical audio models, such as spectral transforms, transmitter parameter mapping, etc. These methods can perform vocal style transitions to some extent but have some limitations, such as unnatural style transition effects, severe distortion problems, etc. To solve these problems, researchers focused on deep learning-based methods, especially generating adversarial networks (GANs). Kaneko et al. proposed a GAN-based method in 2017 to perform voice style transfer through an adversarial relationship between the generator and discriminator [6]. To solve the problem of instability in GAN training, researchers have also tried to introduce Wasserstein GAN (WGAN) [7]. Zhu et al. first proposed CycleGAN for style transfer in the image field [8]. Subsequently, CycleGAN was successfully applied to the voice style transfer task, showing high naturalness [9].

However, the traditional CycleGan network was originally designed for image style transfer, and there are certain limitations in the voice style transfer task, such as the low quality of the generated voice and poor effect.

Therefore, this paper proposes a series of efficiency improvement methods based on the CycleGAN network for voice style transfer tasks. More specifically, the contributions made in this paper are as follows:

  DOI: 10.25236/icceme.2023.013

(1) This paper introduces a second adversarial loss to replace the adversarial loss in the original CycleGAN. By introducing an additional discriminator and calculating an additional opponent loss, the constraints on the generated sound are enforced, thus minimizing the problem of over-smoothing in the statistical model and improving the quality of the generated samples.
(2) The structure of the generator and the discriminator are optimized separately. The generator uses a 2D-1D-2D convolution structure to reduce computational complexity and spectral distortion of generated samples and improve the quality of voice style transfer. The discriminator adopts PatchGAN, which can pay attention to more detailed structural information, thereby generating higher-quality output.
(3) The auxiliary Frame Missing (FIF) self-supervised learning approach is applied to improve the general models in the voice style transfer task. FIF technology sets some images to zero value in the Mel spectrum diagram, which helps the model pay attention to the time-frequency structure of speech during training and improves the model's performance.

The remainder of this paper is as follows. Section 2 summarizes the research background and current status of voice style transfer. In part 3, this paper describes the system's architecture and suggests our improvements. Section 4 describes the experimental setup and results comparison. Section 5 is the conclusion of this paper.

2. Related Work

The evolution of voice style transfer has gradually evolved from the traditional method to the deep learning method. The earliest research can be traced back to the 1980s. In 1988, Abe et al.proposed a parametric method based on vector quantization and spectrum mapping for voice transfer [1].

However, the transition effect is not ideal since the feature space is not continuous. In order to solve this problem, Stylianou et al.proposed a method based on the Gaussian mixture model (GMM) of the sound source spectrum in 1998 [2]. Compared with vector quantization methods, GMM can better match voice features. However, GMM itself is not a one-to-one mapping, which leads to the problem of over-smoothing and over-matching of the converted voice, limiting the technology's development.

With the development of deep learning, researchers began to try to use complex neural network models to model acoustic features. Yao Q et al. proposed a method based on deep neural networks (DNNs) [3], Sn, Lifa et al. used a Short-Term Long-Term Memory (LSTM) network for speech conversion [4], Sato et al. used the high-speed road network for voice conversion [5]. Although these methods have improved regarding naturalness and intelligibility of transitions, the problem of over-smoothing still leads to insufficient naturalness of voice.

As a deep learning model with power generation capabilities, the introduction of GAN has greatly improved the quality and fidelity of the generated speech. As a deep learning network architecture, the GAN is continuously optimized through the interaction between the generator and the discriminator in voice style transfer and can generate dummy data similar to the real data. Compared with traditional methods, GAN-based voice style transfer technology has many advantages, such as no manual data labeling, a more natural conversion effect, and higher conversion efficiency [11].

Kaneko et al. proposed a GAN-based method in 2017 to perform voice style transfer through an adversarial relationship between the generator and discriminator [6]. The generator is responsible for converting the source voice to the target type, while the discriminator evaluates the difference between the generated voice and the actual target type voice. The generator is responsible for converting the source voice to the target type, while the discriminator evaluates the difference between the generated voice and the actual target type voice. In addition, in order to solve the instability problem in GAN training, researchers also tried to introduce WGAN [7]. WGAN improves the training stability by using Wasserstein distance instead of the loss function of traditional GAN.

However, traditional GANs require consistent training data, which is difficult to meet in many situations. To solve this problem, CycleGAN (Cycular Consistency Adversarial Network) came into being. Unlike traditional GANs, CycleGAN does not require concatenated tuples and learns the mapping relationship between the two types through a one-way transformation. By using cycle

consistency constraints, CycleGAN ensures that the output is as consistent as possible with the original input during the reverse transition, making the transitions more natural.

In recent years, CycleGAN has become another important technical direction, which solves the problem of unsupervised image-to-image conversion through cycle consistency loss. Zhu et al. first proposed CycleGAN for style transfer in the image field [8]. Then, CycleGAN was successfully applied to the voice style transfer task, showing high quality and natural transition [9].

3. Proposed Methods

3.1 Second Adversarial Losses

In traditional CycleGAN models, the adversarial loss is used to ensure that the generated sound matches the target pattern [12]. However, the loop consistency loss uses the L1 norm, which can lead to over-smoothing of the generated audio. A second adversarial loss can be introduced to alleviate the problem of over-smoothing in statistical models to solve this problem.

The implementation details of the second adversarial loss strategy include the introduction of an additional discriminator and the calculation of additional adversarial losses [13]. First, the additional D'X and D'Y discriminators for each cycle need to be introduced in the second adversarial loss. These discriminators work with the original DX and DY discriminators but are responsible for the loop-switched audio functions. This additional adversarial loss is used along with the original adversarial loss to apply the adversarial loss to the audio function twice per cycle.

The formula is as follows: (take D'X as an example, D'Y is the same)

$$\mathcal{L}_{adv2}(G_{X\to Y}, G_{Y\to X}, D'_X) = \mathbb{E}_{x\sim P_X(x)}[\log D'_X(x)] + \mathbb{E}_{x\sim P_X(x)}\left[\log\left(1 - D'_X\left(G_{Y\to X}(G_{X\to Y}(x))\right)\right)\right] \quad (1)$$

By introducing a second adversarial loss, the model can better handle the problem of over-smoothing [14]. This method adds an additional adversarial loss to the cycle consistency loss to improve the quality of the generated samples to be closer to real samples, which helps improve the performance of speech conversion tasks.

3.2 Improved Generator: 2D-1D-2D Structure

The generator first converts the input 2D Mel spectrogram into 1D features through convolution operation to improve the 2D-1D-2D structure [15] [16]. The process uses reshaping layers, 1x1 convolution layers, and instance normalization layers to reduce computational complexity while preserving the time-frequency information of the input features. The generator then processes these 1D features using a 1D residual convolution network to better capture long-term dependencies in the Mel spectrogram [17]. After a 1D convolutional network processes the features, the generator recreates the 1D features in a 2D Mel spectrogram using an unwrapping layer to complete the voice-style transmission task. This improved generator structure can improve voice style transfer quality and reduce generated samples' spectral distortion. The schematic diagram of the improved generator *(2D-1D-2D)* is shown in Figure 1.

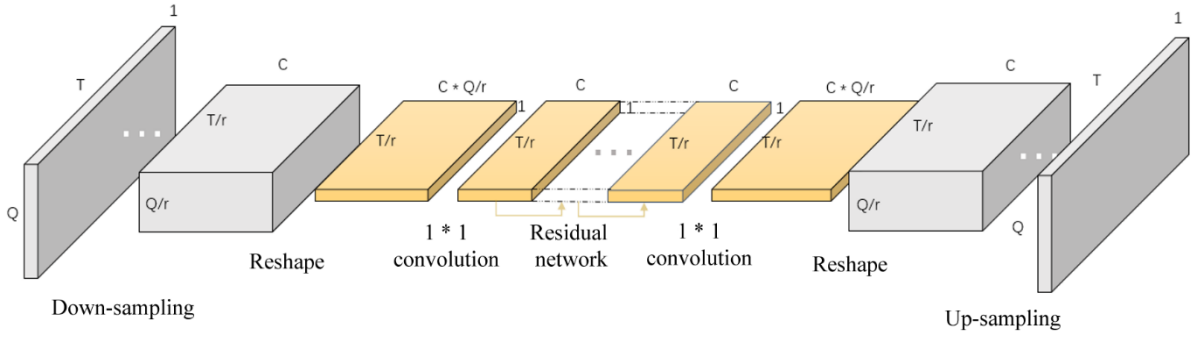

Figure 1 Schematic diagram of the improved generator (2D-1D-2D)

3.3 Improved Discriminator: PatchGAN

PatchGAN is a discriminator that targets local regions of an image or feature map. Unlike traditional full image discriminator tools (such as FullGAN), PatchGAN aims to evaluate whether a local input area belongs to the target category or type. During training, the PatchGAN discriminator evaluates multiple local regions of input features to better capture finer-grained feature differences. This locality encourages the generator to produce features that align with the target's style. In comparison, FullGAN focuses on global consistency while ignoring local structural information. The schematic diagram of the improved discriminator (PatchGAN) is shown in Figure 2 as follows.

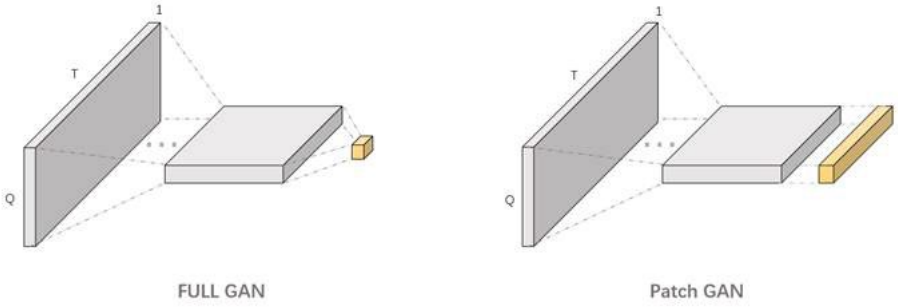

Figure 2 Schematic diagram of the improved discriminator (PatchGAN)

This feature of PatchGAN gives it outstanding performance in styling, image generation, and other general tasks, as it can pay attention to more detailed structural information, resulting in high output quality. Overall, compared with FullGAN, the PatchGAN discriminator improves the performance and output quality of the generalized models by focusing on the local structure of the input feature map.

3.4 The Application of Filled Frame Technology

FIF (Frame Inpainting and Filling) technique is a self-monitoring learning method that improves general patterns in voice type transfer tasks by filling in missing frames to capture time-frequency structure [18] [19]. The principle of the FIF technique is to artificially create missing frames by including a temporary mask in the source mel-spectrogram, which corrects certain frames at zero values. The generator is then trained to fill in those missing frames instead of directly transferring the type. This approach encourages the transmitter to focus on the characteristics of the time-frequency structure, thereby improving the generation quality [20].

The working procedure of the FIF technique is as follows: set a source mel spectrum x, first generate a time mask m of the same size as x, where some regions have zero values (representing the missing frames) and other regions with a value of 1. Subsequently, the m mask is applied to x, producing a mel spectrum xˆ with missing frames. The Gmask generator takes xˆ and m as input and uses m as conditional information to fill in the missing frames, thus generating a y′-filled target mel spectrum. An adversarial loss is used for the constraints to ensure that y' lies in the target domain Y. The schematic diagram of the filling frame technology is shown in Figure 3.

Since there is no parallel data to monitor, the FIF evaluates the stuffing effect due to loss of cycle consistency. Specifically, the Gmask inverse generator reconstructs x" from y', and the cycle consistency loss between the original Mel frequency spectrum x and the reconstructed Mel frequency spectrum x ' is calculated. The generator must extract useful information from the surrounding frames to optimize this loss and fill in the missing frames. This generalization facilitates self-supervised learning of time-frequency structures in mel-spectrograms. Similar effects can be seen in other areas, such as drawing images and filling text.

The main advantage of FIF technology is that it does not require additional data or pre-trained models (such as linguistic information) and can be used as a self-supervised learning method to improve the performance of the style transformation model. And it is enough to double the number of input channels to receive m and xˆ without significantly increasing the model's parameters.

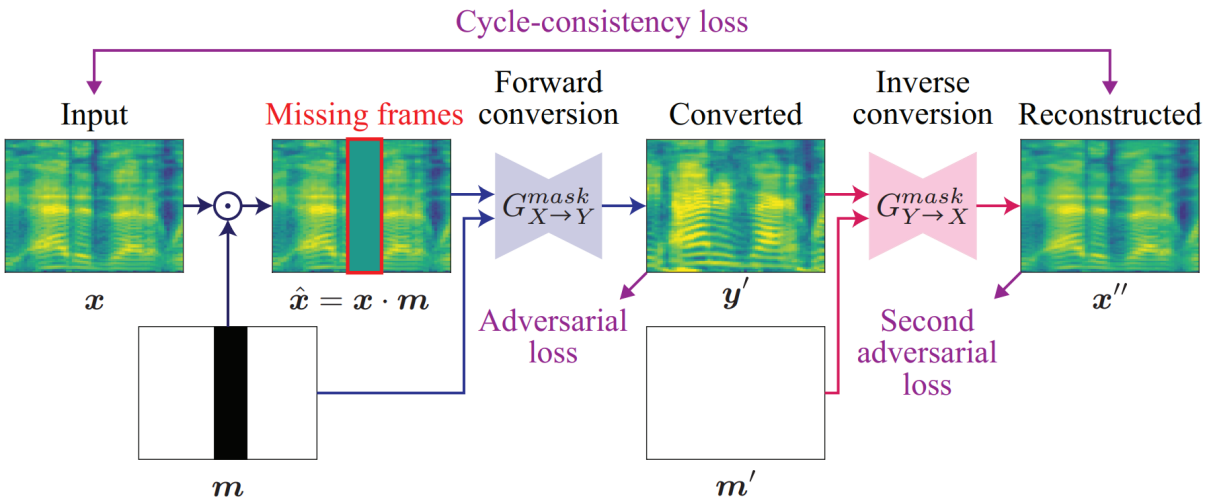

Figure 3 Schematic diagram of filling frame technology

4. Experiments

4.1 Experimental Environment

In this study, to use CycleGAN for voice style transfer, the test environment is configured as follows: The operating system is 64-bit Ubuntu 16.04, the GPU model is NVIDIA RTX 2080 Ti, the display memory is 11 GB, and the memory is 40 GB. In order to be compatible with the pyworld library, the deep learning framework uses the relatively basic PyTorch 1.1, corresponding to Python version 3.7. Using such hardware configuration and software environment, the GPU can be fully utilized to accelerate many matrices and convolution operations during training to gain speed and higher training performance in experience [21].

4.2 Dataset Selection

AISHELL-3 is a large-scale, high-fidelity multi-speaker Mandarin voice database suitable for training various voice systems. After many tests and comparisons, considering the variety of speaker properties, sound fidelity, etc., this paper finally chooses to use the AISHELL-3 database for voice style transfer training. In short, AISHELL-3 has three advantages: (1) High audio fidelity, AISHELL-3 data sets are recorded with a high-fidelity microphone (44.1kHz, 16bit) to ensure sound quality. (2) Larger data scale, the AISHELL-3 dataset contains about 85 hours, and 88035 records, helping to train a more general voice style transfer model. (3) More diverse speaker attributes, the AISHELL-3 dataset includes 176 female speakers and 42 male speakers, making it more suitable for studying voice style transfer between different genders.

From the AISHELL3 database, two female voices, SSB0033 and SSB0145, and two male voices, SSB1863 and SSB0316, were selected and combined into a group for training. Specifically, it contains four transformations in which males and females are combined: female to female, male to male, female to male, and male to female. Among them, there are 500 audios for each timbre, of which 480 are selected as a learning set, and 20 are reserved for testing and experimentation.

4.3 Data Processing

When preprocessing the source audio data, the audio is down-sampled, reducing the sampling rate to 22.05 kHz. Simultaneously, a WORLD analyzer was used to extract 34 Mel cepstrum coefficients (MCEPs), fundamental logarithmic frequency (log F0), and non-periodic indicators (APs) every 5 ms [10]. These features are extracted every 5ms, which can make the extracted features reflect rapid changes in the voice signal while preserving feature resolution. This is important for capturing insights and creating natural voices in voice transfer tasks [22]. Also, to increase the randomness of the training data, a chunk (128 frames) is cut from a randomly selected sentence instead of using the whole sentence directly.

4.4 Hyper Parameter

The training hyperparameters are defined in the following table: num_epochs represents the total number of training epochs, defining the training time of the model. batch_size represents the number of samples contained in each batch. lr represents the learning rate, which is the step size used by the optimizer when updating the model weight [23]. decay_after represents the specified number of epochs after which the learning rate decays so that the model weights can be updated more gently in the later stages of training. num_frames refers to the number of frames per training sample, which will determine the length of the audio clips imported into the model. Table 1 shows the setting of hyperparameters.

Table 1 Setting of hyperparameters

num_epochs	batch_size	lr	decay_after	num_frame
2e4	5	5e-4	1e4	64

4.5 Comparison of Results

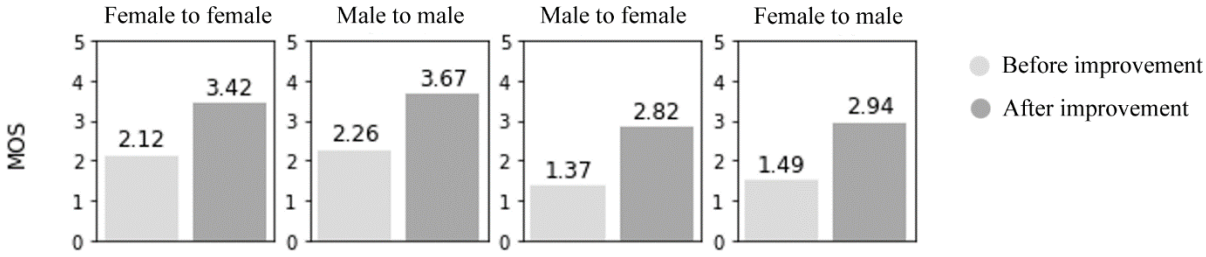

Figure 4 MOS column comparison diagram

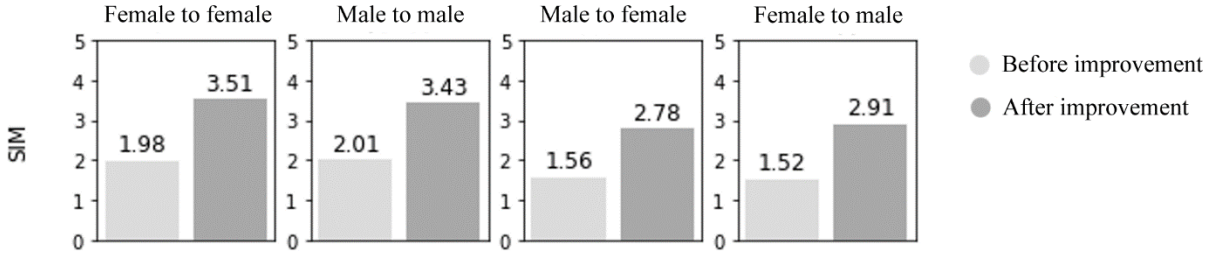

Figure 5 SIM column contrast diagram

Observe the conversion effect between the same gender. The MOS score of female-to-female conversion increased from 2.12 to 3.42, and the SIM score increased from 1.98 to 3.51. The MOS score of male-to-male increased from 2.26 to 3.67, and the SIM score increased from 2.01 to 3.43. These data show that the improved model significantly improves the naturalness and similarity of voices in same-sex transfer situations, making the converted voice closer to the human voice performance. The MOS column comparison diagram and SIM column contrast diagram are shown in Figure 4 and Figure 5, respectively.

Observe the transition effect between the opposite sexes. Male to female MOS score increased from 1.37 to 2.82, and the SIM score increased from 1.56 to 2.91. The MOS score of female-to-male increased from 1.49 to 2.94, and the SIM score increased from 1.52 to 3.12. These results illustrate that the improved model also significantly improves opposite-sex conversion situations.

Although the performance of these two transformations is still slightly lower than the same-sex transfer in terms of naturalness and voice similarity, there has been a significant improvement, laying the basis for further improving the performance of these transformations. Table 2 and Table 3 show the MCD comparison(DB) and MSD Comparison (DB), respectively.

Table 2 MCD comparison (DB)

NO.	Improvement measures			Same-sex transfer		Transgender transfer	
	LOSS	Generator and Discriminator	Missing frame	Female to female	Female to male	Male to female	Female to male
	Tradition CycleGAN(BASELINE)						
0	Single step	2D&FULLGAN	Prevent	7.32	6.97	7.72	7.39
	Improved CycleGAN						
1	**Two-step**	2D&FULLGAN	Prevent	6.98	6.56	7.47	7.14
2	Single step	**2D-1D-2D&PatchGAN**	Prevent	6.61	6.44	7.25	7.06
3	Single step	2D&FULLGAN	**Use**	6.88	6.34	7.35	7.23
4	**Two-step**	2D-1D-2D&PatchGAN	Use	6.47	6.19	6.91	6.98

Table 3 MSD Comparison (DB)

NO.	Improvement measures			Same-sex transfer		Transgender transfer	
	LOSS	Generator and Discriminator	Missing frame	Female to female	Female to male	Male to female	Female to male
	Tradition CycleGAN(BASELINE)						
0	Single step	2D&FULLGAN	Prevent	2.21	2.47	2.71	2.62
	Improved CycleGAN						
1	**Two-step**	2D&FULLGAN	Prevent	1.65	1.57	1.82	1.94
2	Single step	**2D-1D-2D&PatchGAN**	Prevent	1.55	1.62	1.79	1.81
3	Single step	2D&FULLGAN	**Use**	1.39	1.48	1.63	1.70
4	**Two-step**	2D-1D-2D&PatchGAN	Use	1.25	1.33	1.49	1.56

After implementing various improvement measures, we obtained the objective results of the indicator evaluation presented in the table above.

Introducing Second Adversarial Losses: when comparing the traditional CycleGAN (No.0) with the improved CycleGAN (No.1), resulting in a second loss of head-to-head, it can be seen that the MCD and MSD are significantly reduced. This shows that second adversarial losses can improve the quality of the style transfer, making the converted audio closer to the target sound in terms of spectral characteristics and modulation characteristics.

Using 2D-1D-2D constructs and PatchGAN discriminator: comparing the traditional CycleGAN (No.0) with CycleGAN (No. 2) introduces an improved generator and discriminator, using 2D-1D-2D and the PatchGAN discriminator also significantly reduced MCD and MSD, which shows that the method can more effectively capture the time-frequency structure of audio, thereby improving the effect of style transfer.

Introduction of FIF technology (Missing Frame Processing): Compared to the traditional CycleGAN (No. 0) and CycleGAN (No.3) that introduced FIF technology, the use of FIF technology can further reduce MCD and MSD, which shows that the FIF technique helps to understand the time-

frequency structure of the audio and can better retain the structure of the original audio during the transition.

5. Conclusions

This paper proposes three improvement methods based on the CycleGAN network for the voice style transfer task and tests them on the AISHELL-3 dataset. Introducing a second adversarial loss for both same-gender and cross-gender transfers, using 2D-1D-2D structure and PatchGAN discriminator, and using FIF technology can effectively improve the effect of style transfer. Among all the improvement measures, when the three methods are used in combination, MCD and MSD both reach the lowest value and achieve the best voice style transfer effect. The experimental results show that the proposed improved method has significantly improved subjective and objective evaluation indicators compared with the traditional CycleGAN model. In the future, we will try to use other audio features for style transfer and introduce more advanced Generative Adversarial Network (GAN) technology to improve the naturalness of the generated sound and system performance.

Acknowledgments

This paper is supported by the Communication University of China, the Teaching Reform Project "Digital Audio Technology" (JG23104009), and the Communication University of China "College Students Innovation Training Program" project (Program number: S202210033066).

References

[1] Abe, M., Shikano, K., & Kuwabara, H. (1988). Statistical approach to voice conversion. In Proc. ICASSP (Vol. 2, pp. 656-659).

[2] Stylianou, Y., Cappe, O., & Moulines, E. (1998). Continuous probabilistic transform for voice conversion. IEEE Transactions on Speech and Audio Processing, 6(2), 131-142.

[3] Yao, Q., Dong, Y., Che, W., Zhang, Z., & Xu, B. (2014). Deep Neural Networks for Voice Conversion. In Proc. Interspeech (pp. 2257-2261).

[4] Sn, L., & Zen, H. (2016). An End-to-End Approach to Speech Synthesis. In Proc. ICASSP (pp. 4950-4954).

[5] Sato, K., Hwang, S. J., Wang, X., Wang, Y., & Narayanan, S. (2017). Hierarchical Highway Network for Voice Conversion. In Proc. Interspeech (pp. 232)

[6] Kaneko, T., & Kameoka, H. (2017). Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks. In Proc. Interspeech (pp. 959-963).

[7] Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. In Proc. ICML (pp. 214-223).

[8] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proc. ICCV (pp. 2242-2251).

[9] Fang, F., Zhang, J., & Cao, X. (2018). Cycle-consistent Deep Feature Learning for Voice Conversion. In Proc. ICASSP (pp. 5289-5293).

[10] Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. IEICE TRANSACTIONS on Information and Systems, E99.D(7), 1877-1884.

[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

[12] Zhang Xiongwei, Miao Xiaokong, Zeng Xin, et al. Research Status and Prospects of Speech

Conversion Technology [J]. Data Collection and Processing, 2019, 34(5): 753-770.

[13] Pan Xiaoqin, Lu Tianliang, Du Yanhui, Tong Xin. A Review of Speech Synthesis and Conversion Technology Based on Deep Learning [J]. Computer Science, 2021,48(08):200-208.

[14] Ren Qiang. Research and Application of Speech Style Transfer Technology Based on Generative Adversarial Networks [D]. Chongqing University of Technology, 2019.

[15] Li Ting. Research on speech conversion system based on generative confrontation network [D]. Tianjin University, 2019. DOI: 10.27356/d.cnki.gtjdu.2019.001747.

[16] Liu Chang, Wei Weimin, Meng Fanxing, et al. Research Progress on Speech Style Transfer [J]. Computer Science, 2022, 49(6A): 301-308.

[17] Li Yanping, Cao Pan, Zuo Yutao, Zhang Yan, Qian Bo. Speech conversion based on i-vector and variational autoencoder versus generative adversarial network [J]. Acta Automatica Sinica, 2022, 48(07): 1824-1833 .DOI: 10.16383/j.aas.c190733.

[18] Zhang Xiao, Zhang Wei, Wang Wenhao, Wan Yongjing. Speech Conversion Algorithm Based on Multispectral Feature Generative Adversarial Network [J]. Computer Engineering and Science, 2020,42(05):893-901.

[19] Kaneko T, Kameoka H, Tanaka K, et al. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 2019: 6820-6824.

[20] Li Tao. Speech conversion under the condition of non-parallel corpus based on CycleGAN network [D]. Dalian University of Technology, 2018.

[21] Wang Huan, Cai Zhiwei, Xu Xinliang, Zhang Bao, Wu Wenyi. Improved method of audio style transfer based on CycleGAN [J]. Journal of Dalian Minzu University, 2023.

[22] Gao Junfeng, Chen Junguo. Speech conversion based on Style-CycleGAN-VC under non-parallel corpus [J]. Computer Application and Software, 2021,38(09):133-139+159.

[23] Ye Hongliang, Zhu Wanning, Hong Lei. Music style conversion method with human voice based on CQT and Mel spectrum [J]. Computer Science, 2021, 48(S1): 326-330+363.