IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 October 2022, accepted 24 October 2022, date of publication 26 October 2022, date of current version 4 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3217466

RESEARCH ARTICLE

CBFMCycleGAN-VC: Using the Improved MaskCycleGAN-VC to Effectively Predict a Person's Voice After Aging

XIAOQUN ZHOU, LING YU, FANGLIN NIU, AND JUNLIN JIN
School of Electronics and Information Engineering, Liaoning University of Technology, Liaoning 121001, China
Corresponding author: Xiaoqun Zhou (zxq980611@163.com)

ABSTRACT One task of nonparallel speech conversion is to convert the source speaker's speech samples to the target speaker's speech samples, keeping the content unchanged. In view of the advantages of MaskCycleGAN-VC in nonparallel speech conversion, such as small model size and superior performance, our paper uses the basic structure of MaskCycleGAN-VC to improve it and proposes a cyclic boundary method filling in the frame MaskCycleGAN-VC (CBFMCycleGAN-VC) model, which predicts the voice of a person as he ages by using voice samples of his younger self. First, this paper adds speech preprocessing modules, including the Chebyshev low-pass filter and adaptive filter, which increases the robustness of the system. Second, our paper considers the time-domain difference in the weight parameters, which makes it easier to grasp the mapping law of the time-domain structure, with a faster convergence speed. Last, the circular boundary method is introduced to avoid the ringing effect, to enhance the connection between the filled frame and the adjacent frame, and to obtain a better generator. The simulation results show that the CBFMCycleGAN-VC model is more suitable for the speech conversion task of predicting the voices of elderly people, and the convergence speed is faster. The converted voice is also closer to the voice of the target speaker in the time domain and frequency domain. Under the condition that the accuracy rate is similar to that of MaskCycleGAN-VC, the MOS score is 17.5% higher than that of MaskCycleGAN-VC.

INDEX TERMS Nonparallel speech conversion, speech preprocessing, loss, cycle boundary method, CBFMCycleGAN-VC.

I. INTRODUCTION

Speech conversion is a branch of speech synthesis. The purpose of speech conversion is to convert the voice of the source speaker to the voice of another target speaker without changing the language content. Speech conversion greatly complements and expands the function of the text-to-speech (TTS) synthesis system, making it universal and capable of meeting more requirements. The speech conversion of nonparallel corpora is an important hot topic of speech conversion. The research of nonparallel corpora helps speech conversion to achieve the goal in fewer corpora, including the database of missing corpora. Therefore, an investigation of the speech conversion of nonparallel corpora is important.

Early speech conversion is divided into speech analysis, mapping and reconstruction modules. [1] The information contained in speech is divided into language content information and speaker feature information. The analysis module separates the information, and the mapping module maps the speaker feature information to the reconstruction module. The language content information remains unchanged. The speech conversion of the target speaker is realized in the reconstruction module, but its limitation is that a parallel corpus must be utilized. The early models employed dynamic time alignment to align or parallel the corpus and established a more appropriate mapping based on numerous parallel databases. Therefore, how to collect and establish such a large

The associate editor coordinating the review of this manuscript and approving it for publication was Alicia Fornés.

database has become an unavoidable problem that continues to puzzle scholars studying speech conversion. For example, a convolutional neural network, Gaussian mixture model, and Tacotron end-to-end model are designed based on this idea [2]. With the concept of speech conversion, our information about the target speaker should be insufficient or even unknown, so it is more realistic that we obtain nonparallel corpus databases. Therefore, nonparallel data speech conversion is the key direction of future speech conversion. Sun L et al. proposed Phonetic Posteriorgram (PPG)-based, post-speech image alignment technology in [3]. This technology improves dynamic time warping (DTW), connects nonparallel corpora as much as possible, and realizes the processing of nonparallel corpora to a certain extent.

After the introduction of neural networks, the trained model converts speech that is not in the corpus, overcoming the limitations of traditional speech conversion. The WaveNet vocoder and its variants, WaveRNN, WaveGlow, FloWavenet, etc., provide more possibilities for speech conversion. Moreover, due to the correlation between speech conversion and speech synthesis, Park et al. applied the Tacotron end-to-end TTS system to speech conversion, and others applied a similar TTS system based on Transformer [4] to speech conversion and achieved good results. However, these models did not throughly solve the problem of a nonparallel corpus.

Later, the emergence of generative countermeasure networks and their superiority in nonparallel data attracted considerable attention. The generating countermeasure network is composed of a generator and discriminator. After learning how to convert the source picture to the target picture, the generator generates a picture, and then the discriminator identifies whether the generated picture is a real target picture or a false target picture. In this way, the generator and discriminator confront each other and generate a generator that can confuse the discriminator with the fake target picture. CycleGAN was introduced to the field of speech conversion by Rafael ferro et al. in [5]. CycleGAN has shown great advantages in the field of speech conversion where nonparallel training data are lacking. The network is good at style transfer. Speech conversion not only needs to continue to convert the features in the time domain but also attaches importance to the conversion of the features in the frequency domain. The Mel-cepstrum reflects the characteristics of the time-frequency structure of speech, especially the frequency domain, which is more conducive to the generative adversarial network completing the speech conversion task. On the basis of [5] and through continuous improvement, CycleGAN-VC, CycleGAN-VC2 and CycleGAN-VC3 have been successively published. On December 21, Takuhiro Kaneko et al. proposed in [6] that MaskCycleGAN-VC combines the latest filling in frame (FIF) technology with CycleGAN-VC2 and achieves better results.

Based on the overall structure of MaskCycleGAN-VC, our paper adds the speech preprocessing part, and improves the loss of the MaskCycleGAN-VC network. In addition, inspired by the circular boundary method, the FIF technology in [6] is improved, and the algorithm of filling frames with the circular boundary method is proposed. The subjective and objective indicators show that our proposed CBFMCycleGAN-VC model can better achieve the expected objectives of this paper than FastSpeech-VC [7], StarGAN-VC [8], MaskCycleGAN-VC and other models.

The content of this paper is organized as follows: Chapter 2 briefly introduces the composition of the loss of MaskCycleGAN-VC, which is convenient for performing a comparison with the improvement in loss below. The advantages and disadvantages of the FIF technology adopted by MaskCycleGAN-VC are also reviewed in Chapter 2. Chapter 3 introduces the improved CBFMCycleGAN-VC model in this paper. Chapter 4 presents the simulation verification and analysis, and Chapter 5 provides a summary.

II. RELATED WORK

MaskCycleGAN-VC belongs to the category of CycleGAN in speech conversion applications, so it is essentially CycleGAN. The most important aspect of the CycleGAN network is to address losses. In many cases, improper selection of losses often causes failure to converge or poor results. This chapter introduces the composition of losses in MaskCycleGAN-VC and the FIF technology employed by MaskCycleGAN-VC for comparison with the improvement in this paper.

A. LOSS OF MASKCYCLEGAN-VC

CycleGAN-VC2 adds a discriminator based on CycleGAN-VC, adds a second adversarial loss, and improves the loss of CycleGAN-VC. MaskCycleGAN-VC follows the loss of CycleGAN-VC2. The total loss, including antagonism loss, cyclic consistency loss, flag mapping loss and secondary antagonism loss, is applied, as shown in (1).

$$\begin{aligned}\mathcal{L}_{\text{full}} = {} & \mathcal{L}_{\text{adv}}^{\text{X}\to\text{Y}} + \mathcal{L}_{\text{adv}}^{\text{Y}\to\text{X}} + \lambda_{\text{cyc}}\left(\mathcal{L}_{\text{cyc}}^{\text{X}\to\text{Y}\to\text{X}} + \mathcal{L}_{\text{cyc}}^{\text{Y}\to\text{X}\to\text{Y}}\right) \\ & + \lambda_{\text{id}}\left(\mathcal{L}_{\text{id}}^{\text{X}\to\text{Y}} + \text{L}_{\text{id}}^{\text{Y}\to\text{X}}\right) + \mathcal{L}_{\text{adv2}}^{\text{X}\to\text{Y}\to\text{X}} + \mathcal{L}_{\text{adv2}}^{\text{Y}\to\text{X}\to\text{Y}}\end{aligned} \tag{1}$$

where the countermeasure loss $\mathcal{L}_{adv}^{X\to Y}$, $\mathcal{L}_{adv}^{Y\to X}$ is the loss of mutual conversion between the target speaker and the source speaker, which renders the generated speech more authentic. The cyclic consistent loss weighted parameter λ_{cyc} specification is used to prevent the target speaker from simplifying the transformation between the source speaker and the target speaker, and to prevent the generator from directly using the target speaker's voice to deceive the discriminator. If no flag mapping loss is added, the generator will change the image through tone change and gradually deviate from the preset target of speech conversion. The secondary confrontation loss is used to balance the statistical average caused by L1 loss in the confrontation loss. However, this loss does not set the loss part that can be directly affected by the difference in time-domain structure, and is not sensitive to the large difference

in the time-domain structure of two speeches. The update iteration of the loss is small, making it more difficult for the network to learn the conversion of speech features. In the third chapter, this paper proposes an improved method.

B. FILLING IN FRAME

MaskCycleGAN-VC uses the latest FIF filling technology. Compared with CycleGAN-VC2, MaskCycleGAN-VC with FIF filling technology more easily captures the time-frequency structure and optimizes the results. Although the Mel-spectrum has been introduced to CycleGAN-VC3, the CycleGAN-VC3 model is large due to the addition of an additional time-frequency adaptive normalization module. In contrast, the MaskCycleGAN-VC model has achieved better performance than the CycleGAN-VC2 model and CycleGAN-VC3 model while it is smaller than the CycleGAN-VC3 model.

FIF filling technology is a technology to fill the vacant frame by surrounding frames, which makes it easier to address conversions with gaps in the time-frequency structure. The equation of this technology is used as follows:

$$\hat{X} = X * \mathcal{M} \tag{2}$$

After filling the vacant frame, it is sent to the subsequent network for circulation and obtains better results. Although FIF achieves better results than the time-frequency adaptive normalization (TFAN) module, the missing frames filled by FIF are not continuous with the frames at both sides, and the matrix is random, which cannot help CycleGAN-VC learn.

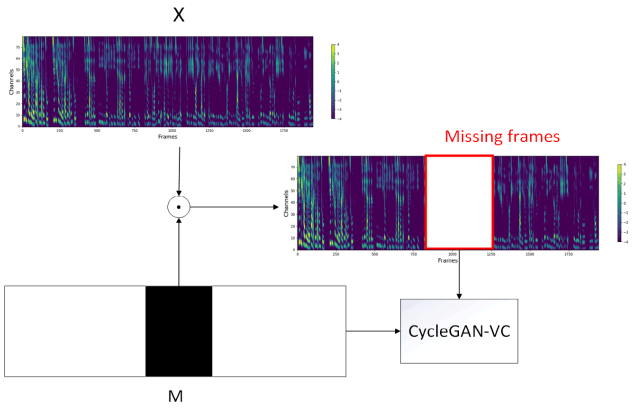


FIGURE 1. Schematic of filling in frames. After X is multiplied by matrix M, the missing frames are filled and entered into CycleGAN-VC as input.

III. CBFMCycleGAN-VC

As shown in Fig. 2, The structure of the CBFMCycleGAN-VC model is proposed. The overall architecture follows the structure of MaskCycleGAN-VC. In the dataset preprocessing stage, a low-pass filter and LMS adaptive noise reduction algorithm are added to improve the robustness of the system. In this stage, the problem that most of the nonactively matched speech database samples contain Gaussian

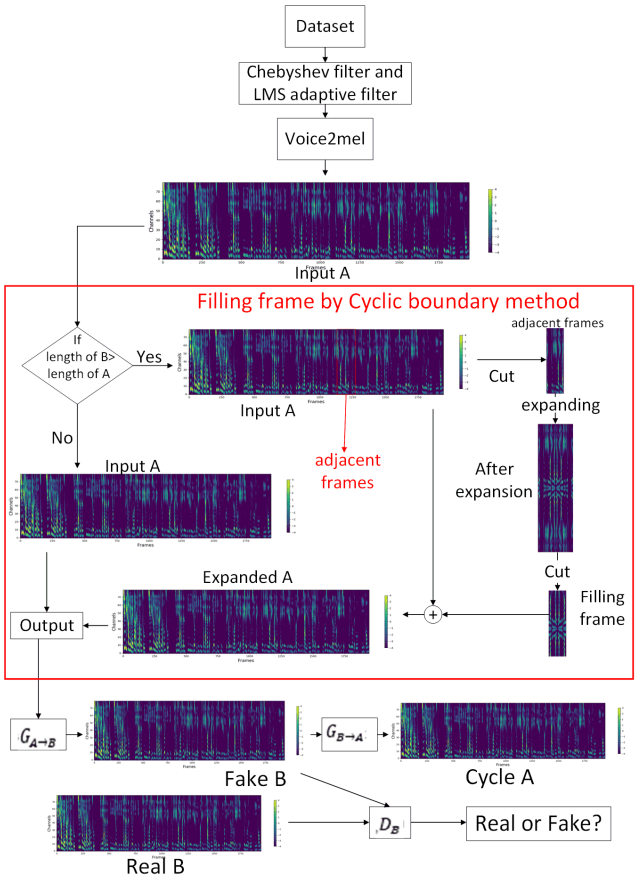


FIGURE 2. Structure diagram of the CBFMCycleGAN-VC model. The red box represents the process of filling the frame via the loop boundary method. The loss is reflected in the generator and discriminator, which is not shown in the Fig.2.

white noise of different degrees in the actual environment is addressed. According to the task characteristics of the speech conversion target proposed in this paper, the loss function is improved, so that the loss function can better grasp the mapping on the time-domain structure. In addition, this paper also uses the cyclic boundary method to improve the FIF, which can effectively supplement the information of the missing frame. The filling frame expanded by the cyclic boundary method is more consistent with the continuous structure in the time domain and is more conducive to the generation of the learning of the countermeasure network.

After the dataset is processed by the Chebyshev filter and LMS adaptive filter, relatively pure speech is obtained, which is converted to the Mel-spectrum through windowing and a fast Fourier transform, i.e., input A. In this paper, the source speaker's voice in the database is named A, and the target speaker's voice is named B. Since the speech time domain length of the source speaker and that of the target speaker differ, it is determined whether the length of B in the database is longer than the input A. If so, extended A is obtained through the cyclic boundary method to reduce the difference between the two, and then generator $G_{A\to B}$ is converted to

fake B, which is then sent to generator $G_{B\to A}$ to generate cycle A. At this time, the obtained cycle A should be similar to input A or extended A, but it is generated by two generators. At this time, input A or extended A, fake B and cycle A will be sent to discriminators D_A and D_B. The second term Through constant confrontation, the generated fake B will become increasingly similar to B, and the discriminator will be increasingly difficult to identify. The above is the whole process of the cycle. Because A and B are nonparallel speech, we also participate in the loop of B's speech in the same way. The discriminator and generator are unchanged. After a continuous loop, we obtain generator $G_{A\to B}$, which can cheat the discriminator. This generator is the final requirement of the speech conversion task. When inputting the young voice A, generator $G_{A\to B}$ generates an output of sound in old age.

A. SPEECH BATCH PREPROCESSING

The purpose of this pretreatment is to remove the Gaussian white noise from the samples obtained from the network and to enhance the practicability of the samples. The reason for using this algorithm is that most of the non-actively-matched speech database samples contain Gaussian white noise, but the size of the noise is different. The manual processing method is obviously too cumbersome. Moreover, after setting this preprocessing noise reduction module, the collection requirements of the speech database are reduced and the practicability of the system is enhanced. In this paper, the LMS algorithm is used to batch process the collected speech materials. The LMS algorithm equation is used as follows:

$$\mathcal{W}(n+1) = \mathcal{W}(n) + 2\mu X(n)e(n) \tag{3}$$

B. CYCLIC BOUNDARY METHOD FILLING FRAME

Although the FIF technology adopted by MaskCycleGAN-VC in [3] has completed the task of filling, the filling frame and the frames at both ends are not continuous, and the truncation between the two is quite obvious, which easily causes a ringing effect. Moreover, because the matrix is random, the filled frame cannot sufficiently help CycleGAN-VC learn. Therefore, the CBFMCycleGAN-VC model proposed in this paper adopts the circular boundary method, which can effectively suppress the ringing effect caused by boundary truncation while expanding the image. The circular boundary method extends the observation image in a reflection symmetric manner, that is, the original image is extended in a symmetrical downward, rightward, and downward right manner. In this paper, the young speech is segmented according to syllables, and the adjacent frames of the syllables whose time domain structure changes by more than 170% of themselves are extracted, then horizontally, vertically and diagonally expanded, and spliced to obtain expanded frames. At this time, the extended frame does not lose too much information and has coherence. We cut out the filling frame from the center and add it to the original young speech to avoid the ringing effect caused by the truncation and to obtain more filler speech for rich padding frames.

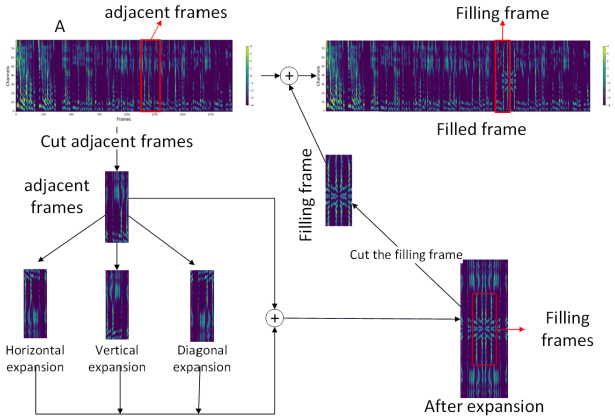


FIGURE 3. Detailed flow chart of the filling frame with the circular boundary method.

As shown in Fig. 3, the method used to supplement the missing frame in this paper is to add the adjacent frame A_{n-1} to carry out the expansion by using the cyclic boundary method and intercept the frame from the center of the expanded image to fill in. The filled frame is more consistent with the original frame on both sides, avoids the ringing effect, and contains more information than the filled frame of FIF, which is more conducive to generating the counter network for learning to obtain a better generator.

C. LOSS

Considering that the task of speech conversion is to convert a person's voice to the aging voice, although the two are different in the frequency domain, the difference is not substantial. However, the difference is larger in the time domain. Therefore, in the design of loss, this paper prefers to supplement the loss with the actual time domain gap, so this paper adds the consideration of the time domain to the weight parameter. The loss proposed in this paper is represented by (4):

$$\begin{aligned} L_{full} &= L_{adl}^{A\to B} + L_{adl}^{B\to A} \\ &\quad + \varpi_{ccl}^{A\to B} \cdot L_{ccl}^{A\to B\to A} + \varpi_{ccl}^{B\to A} \cdot L_{ccl}^{B\to A\to B} \\ &\quad + \lambda_{idl} L_{idl}^{A\to B} + \lambda_{idl} L_{idl}^{B\to A} + L_{adl}^{A\to B\to A} + L_{adl}^{B\to A\to B} \end{aligned} \tag{4}$$

First, the first term of (4) is the countermeasure loss, and the countermeasure loss is the generator $G_{A\to B}$, which is the loss corresponding to the generation of old voice B from young voice A. The L1 loss is used for the cycle consistency and the identity mapping but not for the adversarial loss of Equation (5).

$$\begin{aligned} L_{adl}^{A\to B} &= \mathbb{E}_{B\sim P(B)}[logD_B(B)] + \\ &\quad \mathbb{E}_A \sim P(A)[logD_B(1 - D_B(G_{A\to B}(A)))] \end{aligned} \tag{5}$$

The discriminator discriminates by maximizing the loss, and the generator generates more deceptive speech by minimizing the loss. The two fight and produce better results.

$\mathbb{E}_{B\sim P(B)}$ represents the expected value of B in a given distribution P (b), and the following similar expressions are the same and will not be repeated. Similarly, $L_{adl}^{B\to A}$ is represented by $G_{B\to A}$, which is the corresponding loss when generated speech B is converted to speech A.

The second term of (4) is the cyclic consistency loss, which is the loss required after A generates B and then generates A, to prevent the mapping from being too simple to obtain an ideal effect in the process of generating A from young voice A to old voice B and then generating A. The equation of cyclic consistency loss is used as follows:

$$\varpi_{ccl}^{A\to B}\cdot L_{ccl}^{A\to B\to A}=\varpi_{ccl}^{A\to B}\cdot \mathbb{E}_{A\sim P(A)}[\|G_{B\to A}\left(G_{A\to B}\left(A\right)\right)-A\|_1 \tag{6}$$

where $\|\cdot\|_1$ represents the L1 norm and ϖ_{ccl} is a weight parameter used to increase its sensitivity to the time domain structure. The calculation equation of ϖ_{ccl} is used as follows:

$$\varpi_{ccl}^{A\to B}=\frac{t_A-\frac{1}{N}\sum_{P(B)}t_B}{\frac{1}{N}\sum_{P(B)}\sqrt{\left(t_n-\frac{1}{N}\sum_{P(B)}t_B\right)^2}} \tag{7}$$

where t_B is the time domain length of the old voice, $P(B)$ is the given distribution of old voice B in the time domain structure, and the distribution method is the same as that used to calculate the old voice expectation in the adversarial loss and cycle consistency loss. N is the total number of data in the dataset, where we set N = 81; $n\in[1,2,\ldots,81]$. $\varpi_{ccl}^{B\to A}$ will also be used as the weight parameter of $L_{ccl}^{B\to A\to B}$.

The third item in Equation (4) is identity loss, which prevents the generator from mapping through simple tone change. Notably, this is not the mapping that the voice conversion task wants the generator to generate. The calculation equation of identity loss is used as follows:

$$\lambda_{idl}L_{idl}^{A\to B}=\mathbb{E}_{B\sim P(B)}\left[\|G_{A\to B}\left(B\right)-B\|_1\right] \tag{8}$$

where λ_{idl} is the weight parameter of identity loss; this parameter reflects the importance this paper attaches to the gap in the time domain structure. The gap between the young voice and the old voice in the time domain will directly affect the update speed of the loss. The larger the gap is, the faster the update speed.

The second countermeasure loss is used to balance the loss of the L1 norm to avoid the generation of countermeasure networks that cannot converge. The equation of the second countermeasure loss is used as follows:

$$L_{adl}^{A\to B\to A}=\mathbb{E}_{A\sim P(A)}[\log D'_A(A)]+\mathbb{E}_{A\sim P(A)}\left[\log\left(1-D'_A\left(G_{B\to A}\left(G_{A\to B}\left(A\right)\right)\right)\right)\right] \tag{9}$$

where D'_A is A obtained after a cycle, which is different from the original speech A. Similarly, $L_{adl}^{B\to A\to B}$ is also used to balance the loss of the L1 norm.

IV. EXPERIMENT

A. DATA COLLECTION AND PREPROCESSING

This article collected several videos about Trump's interviews when he was young and old and the video of the president's swearing in speech on the internet, and intercepted the speech fragments, for a total of 81 voice records when he was young and 81 voice records when he was old. As the training database used in this article, the voice records in the database are nonparallel voices, with a length of 2.8-7.9 seconds.

Since the TV program from which the speech segment is intercepted is a program many years ago, the speech contains low-frequency Gaussian white noise by drawing the time-domain diagram and frequency-domain diagram. Considering that speech is intercepted from different videos, this paper uses the aforementioned preprocessing module to process speech before the speech conversion system. The preprocessing module designed in this paper basically completely eliminates the Gaussian white noise in the frequency part of 200 Hz-20000 Hz and suppresses 90% of the Gaussian white noise in the frequency part of 20 Hz-200 Hz, which basically realizes the data processing requirements of this paper.

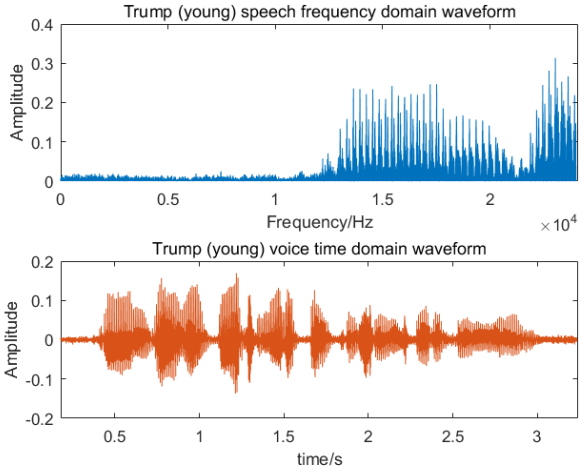
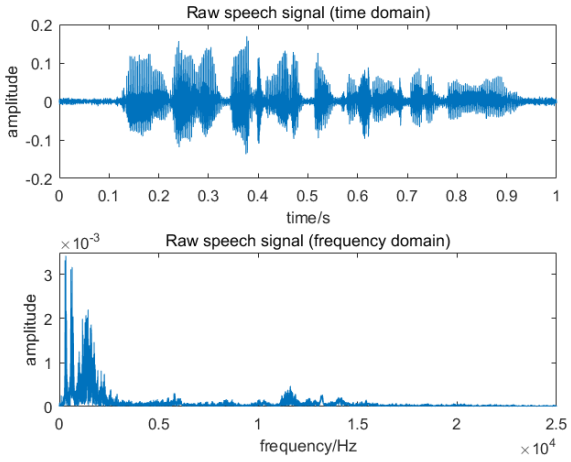


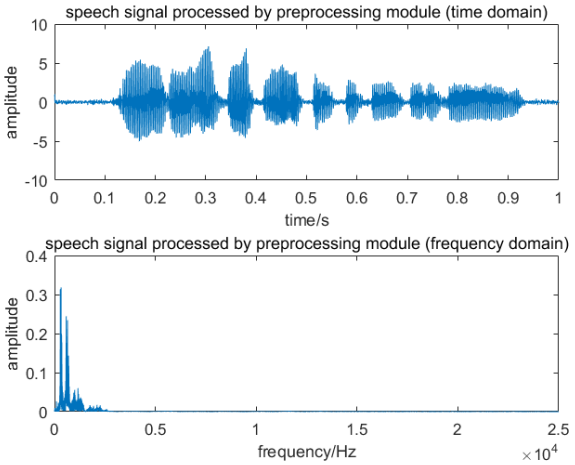
FIGURE 4. Trump (young) voice time domain waveform.

According to the analysis of the speech signal by the Fourier transform, a Chebyshev I low-pass filter is designed. The passband frequency of the filter is set to 2.5 kHz, and the stopband frequency is set to 25 kHz. The LMS needs a reference signal $\boldsymbol{D}(\boldsymbol{n})$, so this paper adds a low-pass filter to the collected speech for preliminary processing and selects it as a reference signal. The order of the LMS is set to 50 orders, and the step size is 0.000008. As shown in Fig. 5, both of them are very effective for noise reduction processing of speech signals in the dataset.

After noise reduction processing, the dataset is subjected to fast Fourier transform using a window size of 1024 with a length of 40 milliseconds and a jump length of 10 milliseconds to obtain an 80-dimensional Mel-spectrum.



(a)



(b)

FIGURE 5. The time domain and frequency domain diagram of the source speech and the speech after LMS noise reduction are the time domain diagram of the unfiltered speech, the time domain diagram of the filtered speech, the frequency domain diagram of the unfiltered speech, and the frequency domain diagram of the filtered speech from top to bottom. (a) The time domain and frequency domain diagram of the raw speech signal. (b) The time domain and frequency domain diagram of speech signal processed by preprocessing module.

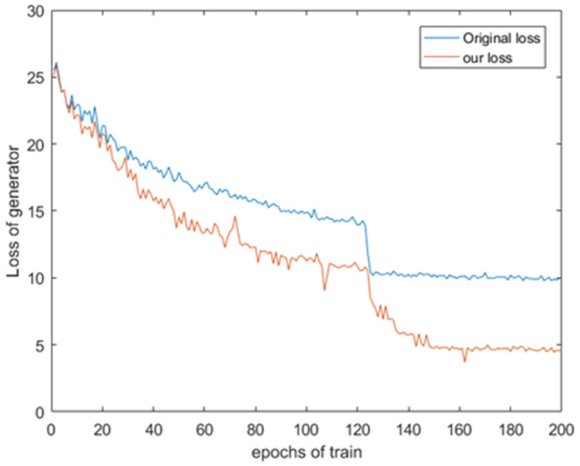


FIGURE 6. $G_{A\to B}$ loss comparison chart.

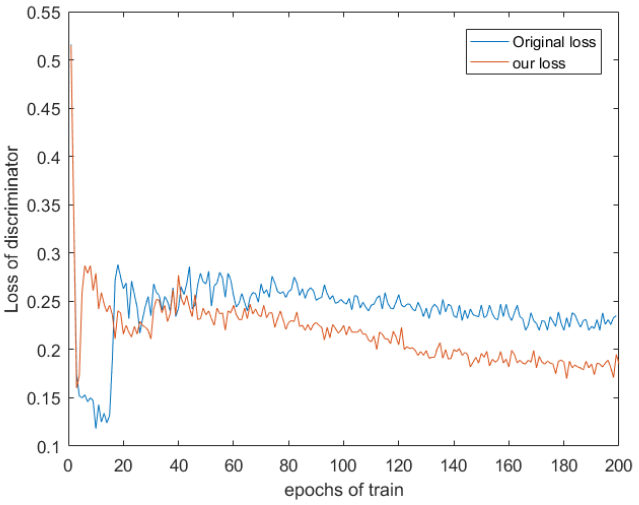


FIGURE 7. $G_{A\to B}$ loss comparison chart.

B. TRAINING SETTINGS

This paper verifies whether the improved loss is effective in the CBFMCycleGAN-VC model by pretraining the model with different losses for 200 rounds. The original loss in Fig. 6 and Fig. 7 means that the loss of MaskCycleGAN-VC in [3] is substituted into the CBFMCycleGAN-VC model for the experiment, and the other line is obtained by our proposed loss test in CBFMCycleGAN-VC. The two are compared, and the drawn loss diagram is shown in Fig. 6 and Fig. 7.

As shown in Fig. 6 and Fig. 7, the discriminator D_A basically completed the convergence in the 146th round, which was 14 rounds earlier than the loss of MaskCycleGAN-VC in the 160th round, and the generator $G_{A\to B}$ converged to 5 in the 150th round. In comparison, MaskCycleGAN-VC did not converge to 10 until the 200th round. Notably, the loss of the model proposed in this paper can achieve a faster convergence speed in the task of transforming young voices into old voices, which is more conducive to fast training and saves time.

The generator of the CBFMCycleGAN-VC model is composed of a 2-1-2d CNN, and the discriminator is PatchGAN. Compared with MaskCycleGAN-VC, CBFMCycleGAN-VC does not need to receive the M matrix mentioned in (2), so it does not need to expand the input channel. In the training, the Adam optimizer training network is used for 200K iterations, and the initial values of the discriminator and generator are set to 0.5 and 25, respectively. The learning rates of the discriminator and the generator are set to 0.0001 and 0.0002, respectively, the batch size is set to 4, and λ_{idl} is set to 5.

C. TEST RESULTS

In this paper, the speech generated by the CBFMCycleGAN-VC model is compared with the speech generated by MaskCycleGAN-VC in many ways. First, this paper selects several words to form multiple phrases; forms the speech

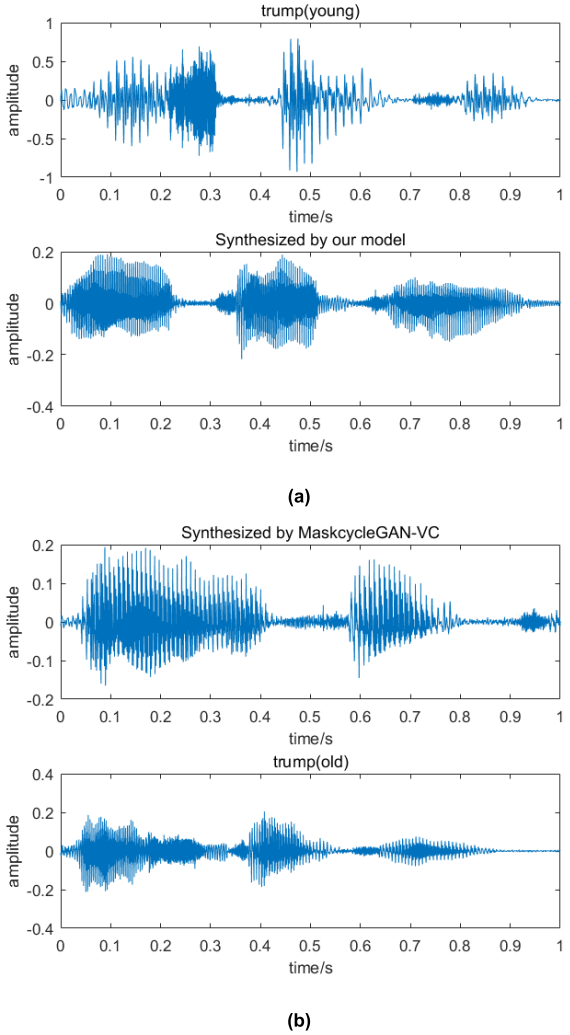


FIGURE 8. Comparative experiment (time domain). (a) Time domain of trump (young) and speech signal synthesized by our model. (b) Time domain of trump (old) and speech signal synthesized by MaskcycleGAN-VC.

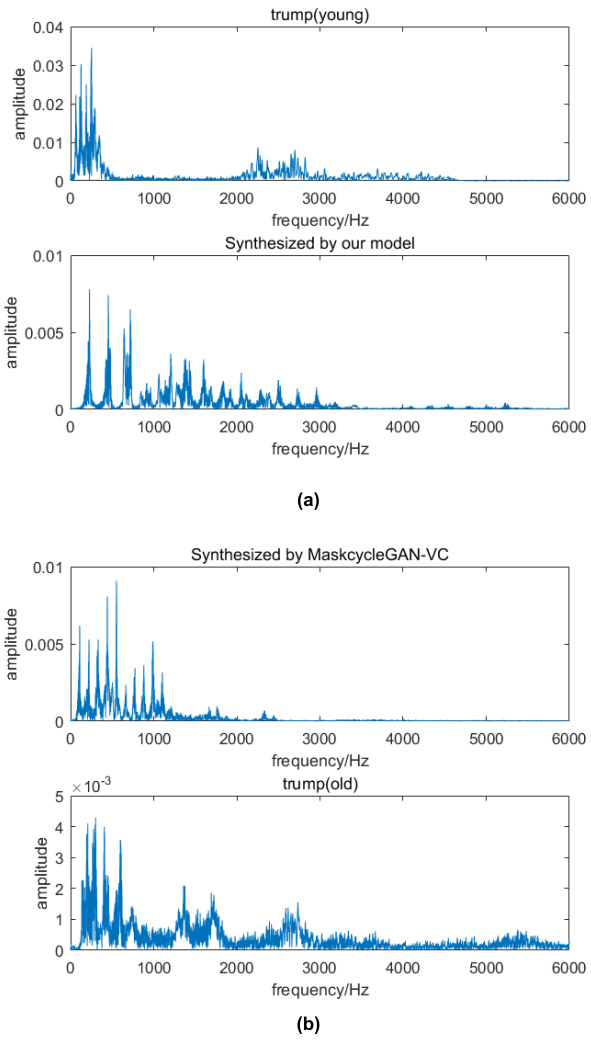


FIGURE 9. Comparison test (frequency domain). (a) Frequency domain of trump (young) and speech signal synthesized by our model. (b) Frequency domain of trump (old) and speech signal synthesized by MaskcycleGAN-VC.

of young people, the speech of old people, the speech synthesized by MaskCycleGAN-VC, and the speech synthesized by the improved model in this paper; and observes the similarity between their time domain and the frequency domain to analyze whether the model proposed in this paper is better than the MaskCycleGAN-VC model. Fig. 8 and Fig. 9 are comparative images of a phrase composed of the words "this" and "country".

As shown in Fig. 9, compared with MaskCycleGAN-VC, the speech synthesized by CBFMCycleGAN-VC is more consistent with not only the time domain characteristics of the target speaker but also old Trump's speaking habits in the time domain, and the transformation in the frequency domain is also closer to the target, Fig. 6 and Fig. 7 shows that MaskCycleGAN-VC has obviously not completed the transformation of this part, especially between 1000 Hz and 2000 Hz. However, the improved model in this paper pays attention to and learns this mapping, and produces a voice more like old Trump.

TABLE 1. Accuracy and MOS for the different VC systems.

Speech Signal Class	Acc%	MOS	
		Naturalness	Similarity
Target(old)	100	4.96±0.04	4.98±0.05
CycleGAN-VC3	97.883	2.44±0.26	2.48±0.12
MaskCycleGAN-VC	**98.412**	2.56±0.28	2.86±0.34
FastSpeech-VC	97.707	2.34±0.35	2.41±0.45
StarGAN-VC	96.825	2.41±0.46	2.23±0.24
Our model	98.059	**3.01±0.38**	**3.11±0.43**

Table 1 shows that various models are equal in accuracy. Note that severely distorted words are also listed as incorrect words. The model proposed in this paper has achieved good results in terms of accuracy and MOS

D. SUBJECTIVE INDEX

This paper also compares the similarity, accuracy and naturalness of speech produced by several different models. The comparison results are shown in Table 1.

TABLE 2. Evaluation form of four people.

Speaker	Speech speed	Accuracy rate	Similarity
MEN A	4.3	4.2	3.4
MEN B	3.8	4.3	3.2
WOMEN A	3.5	3.5	2.7
WOMEN B	3.9	3.2	2.4

In this paper, $G_{A\to B}$, which represents the voices of the four old people, were inversely predicted, and they were asked to judge whether their voices were similar to their young voices. The model proposed in this paper has a good score in the prediction of speech speed, but the accuracy and similarity are poor. Moreover, the prediction of our model for men is better than that for women

In addition to MOS scores, this paper also conducted a reverse verification by collecting the voices of four old people both when they were young and now, removing $G_{A\to B}$ to restore the voices of the four old people to the voices of the young people, and proposing some questions. The full score is 5 points. The evaluation is shown in Table 2.

V. CONCLUSION

The purpose of speech conversion in this paper is to predict the voice of a person when he or she ages. Generally, the speech speed of young people is generally faster, and the speech speed of old people is generally slower and more turbid. The time difference between the two may be greater than that of male and female conversion. Therefore, this paper updates the loss function with the time-domain length improvement weight parameter. This paper also proposes a new additional module to improve the performance of CycleGAN, making it more consistent with the experimental purpose of this paper. Fig. 8, Fig. 9, Table 1 and Table 2 show that the model basically achieved the goal of predicting the sounds of the elderly through the sounds of young people. The CBFMCycleGAN-VC model obtained a score of 98.059% in the accuracy rate, which is relatively stable and surpasses most models. The MOS naturalness score obtained 3.01 ± 0.38, and the similarity score obtained 3.11 ± 0.43, both of which are in the leading position. The model also reversely passed $G_{B\to A}$, which reversely verified the effectiveness of the model. The model is successful in the transformation of men and slightly poor in the transformation of women.

REFERENCES

[1] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 132–157, Nov. 2021, doi: 10.1109/TASLP.2020.3038524.

[2] R. Levy-Leshem and R. Giryes, "Taco-VC: A single speaker tacotron based voice conversion with limited data," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Amsterdam, The Netherlands, Jan. 2021, pp. 585–590, doi: 10.23919/Eusipco47968.2020.9287448.

[3] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Seattle, WA, USA, Jul. 2016, pp. 1–6.

[4] C. Fu, C. Liu, C. T. Ishi, and H. Ishiguro, "CycleTransGAN-EVC: A CycleGAN-based emotional voice conversion model with transformer," 2021, *arXiv:2111.15159*.

[5] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 2100–2104.

[6] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Maskcyclegan-VC: Learning non-parallel voice conversion with filling in frames," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 5919–5923.

[7] S. Zhao, H. Wang, T. H. Nguyen, and B. Ma, "Towards natural and controllable cross-lingual voice conversion based on neural TTS model and phonetic posteriorgram," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 5969–5973.

[8] M. Baas and H. Kamper, "StarGAN-ZSVC: Towards zero-shot voice conversion in low-resource contexts," in *Proc. Southern Afr. Conf. Artif. Intell. Res.* APretoria, South Africa: Springer, Feb. 2021, pp. 69–84, doi: 10.1007/978-3-030-66151-9_5.

[9] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-VC2: Improved cyclegan-based non-parallel voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 6820–6824.

[10] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC3: Examining and improving CycleGAN-VCs for mel-spectrogram conversion," in *Proc. Interspeech*, Shanghai, China, Oct. 2020, pp. 2017–2021.

[11] M. Patel, M. Parmar, S. Doshi, N. J. Shah, and H. A. Patil, "Novel adaptive generative adversarial network for voice conversion," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*,), Lanzhou, China, Nov. 2019, pp. 1273–1281.

[12] Y.-H. Chen, D.-Y. Wu, T.-H. Wu, and H.-Y. Lee, "Again-VC: A one-shot voice conversion using activation guidance and adaptive instance normalization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 5954–5958.

[13] X. Kang, H. Huang, Y. Hu, and Z. Huang, "Connectionist temporal classification loss for vector quantized variational autoencoder in zero-shot voice conversion," *Digit. Signal Process.*, vol. 116, Sep. 2021, Art. no. 103110, doi: 10.1016/j.dsp.2021.103110.

[14] X. Zhang, J. Wang, N. Cheng, E. Xiao, and J. Xiao, "Cyclegean: Cycle generative enhanced adversarial network for voice conversion," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Cartagena, CO, USA, Dec. 2021, pp. 930–937.

[15] Y. Liu, A. Chen, H. Shi, S. Huang, W. Zheng, Z. Liu, Q. Zhang, and X. Yang, "CT synthesis from MRI using multi-cycle GAN for head-and-neck radiation therapy," *Computerized Med. Imag. Graph.*, vol. 91, Jul. 2021, Art. no. 101953, doi: 10.1016/j.compmedimag.2021.101953.

[16] B. Sisman, M. Zhang, M. Dong, and H. Li, "On the study of generative adversarial networks for cross-lingual voice conversion," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Singapore, Dec. 2019, pp. 144–151.

[17] R. Elkadiri, C. Manche, M. Sultan, A. Al-Dousari, S. Uddin, K. Chouinard, and A. Z. Abotalib, "Development of a coupled spatiotemporal algal Bloom model for coastal areas: A remote sensing and data mining-based approach," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 11, pp. 5159–5171, Nov. 2016, doi: 10.1109/JSTARS.2016.2555898.

[18] C. Wang and Y.-B. Yu, "CycleGAN-VC-GP: Improved CycleGAN-based non-parallel voice conversion," in *Proc. IEEE 20th Int. Conf. Commun. Technol. (ICCT)*, Nanjing, China, Oct. 2020, pp. 1281–1284.

[19] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Commun.*, vol. 88, pp. 65–82, Apr. 2017, doi: https://doi.org/10.1016/j.specom.2017.01.008.

[20] Z. Du, "Spectrum and prosody conversion for cross-lingual voice conversion with cyclegan," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Auckland, New Zealand, Dec. 2020, pp. 507–513.

[21] J.-W. Kim, H.-Y. Jung, and M. Lee, "Vocoder-free End-to-End voice conversion with transformer network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Glasgow, U.K., Jul. 2020, pp. 1–8.

[22] S. Shi, J. Shao, Y. Hao, Y. Du, and J. Fan, "U-GAT-VC: Unsupervised generative attentional networks for non-parallel voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 7017–7021.

[23] R. Ferro, N. Obin, and A. Roebel, "CycleGAN voice conversion of spectral envelopes using adversarial weights," in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Amsterdam, The Netherlands, Jan. 2021, pp. 406–410.

[24] Y. Alaa, M. Alfonse, and M. M. Aref, "A survey on generative adversarial networks based models for Many-to-many non-parallel voice conversion," in *Proc. 5th Int. Conf. Comput. Informat. (ICCI)*, New Cairo, Egypt, Mar. 2022, pp. 221–226.
[25] C. Wen, T. Guo, X. Tan, R. Yan, S. Zhou, C. Xie, W. Zou, and X. Li, "Time domain adversarial voice conversion for ADD 2022," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 9221–9225.

XIAOQUN ZHOU received the B.S. degree from Qufu Normal University, Jinzhou, China, in 2020. He is currently pursuing the M.S. degree with the School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou. His research interests include natural language processing and deep learning.

LING YU received the B.S. degree in applied electronic technology from the Liaoning Institute of Technology, Jinzhou, China, in 2002, the M.S. degree in communication and information system from the Liaoning University of Technology, Jinzhou, in 2008, and the Ph.D. degree in signal and information processing from the Dalian University of Technology, Dalian, China, in 2017. She was with the Liaoning University of Technology, in 2002, where she is currently an Associate Professor at the School of Electronics and Information Engineering. Her research interests include non-Gaussian signal processing and time delay estimation.

FANGLIN NIU received the B.S. and Ph.D. degrees from the Dalian University of Technology, Dalian, China, in 1996 and 2015, respectively. She is currently an Associate Professor at the School of Electronics and Information Engineering, Liaoning University of Technology. Her research interests include information theory, channel coding, fountain codes, and wireless communication technology.

JUNLIN JIN received the B.S. degree from the Suqian College, Jiangsu, Xuzhou, China, in 2020. He is currently pursuing the M.S. degree with the School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou, China. His research interests include natural language processing and deep learning.