Emotion Speech Synthesis Method Based on Multi-Channel Time–Frequency Domain Generative Adversarial Networks (MC-TFD GANs) and Mixup

Ning Jia¹ · Chunjun Zheng¹

Received: 2 June 2021 / Accepted: 12 August 2021 / Published online: 24 August 2021
© The Author(s) 2021

Abstract

As one of the most challenging and promising topics in speech field, emotion speech synthesis is a hot topic in current research. At present, the emotion expression ability, synthesis speed and robustness of synthetic speech need to be improved. Cycle-consistent Adversarial Networks (CycleGAN) provides a two-way breakthrough in the transformation of emotional corpus information. But there is still a gap between the real target and the synthesis speech. In order to narrow this gap, we propose an emotion speech synthesis method combining multi-channel Time–frequency Domain Generative Adversarial Networks (MC-TFD GANs) and Mixup. It includes three stages: multichannel Time–frequency Domain GANs (MC-TFD GANs), loss estimation based on Mixup and effective emotion region stacking based on Mixup. Among them, the gating unit GTLU (gated tanh linear units) and the image expression method of speech saliency region are designed. It combines the Time–frequency Domain MaskCycleGAN based on improved GTLU and the time-domain CycleGAN based on saliency region to form the multi-channel GAN in the first stage. Based on Mixup method, the calculation method of loss and the aggravation degree of emotion region are designed. Compared with several popular speech synthesis methods, the comparative experiments were carried out on the interactive emotional dynamic motion capture (IEMOCAP) corpus. The bi-directional three-layer long short-term memory (LSTM) model was used as the verification model. The experimental results showed that the mean opinion score (MOS) and the unweighted accuracy (UA) of the speech generated by the synthesis method were improved, and the improvements were 4% and 2.7%, respectively. The current model was superior to the existing GANs model in subjective evaluation and objective experiments, ensure that the speech generated by this model had higher reliability, better fluency and emotional expression ability.

Keywords Multi-channel time–frequency domain generative adversarial networks · MaskCycleGAN · Mixup · Emotion speech synthesis · Gated tanh linear units · Mel-cepstrum · Multi-channel technology

1 Introduction

As one of the popular fields of speech research, speech emotion interaction has potential application value in health care [1], call center [2] and human–computer interaction based on natural speech [3]. Although the research of deep learning technology has made great progress, the performance of the most advanced speech emotion interaction system is not well. Lack of data is one of the main bottlenecks in this

field [4]. Compared with other speech tasks, the available speech emotion recognition corpus is relatively small, which directly leads to the disaster of speech feature dimension and limits the performance of emotion recognition [5]. One of the common methods to solve the problem of data shortage is dimension reduction technology [6]. However, we cannot guarantee that the extracted features provide the best performance in emotion recognition [7].

Another factor that affects the speech emotion interaction system is the ability of speech synthesis. Judging from the current development of speech synthesis technology, the clarity and intelligibility of synthetic speech have been greatly improved. Moreover, the naturalness of the synthesized speech is greatly improved. However, in the traditional speech synthesis, people often simply solve the problem of

✉ Ning Jia
jianing@neusoft.edu.cn

¹ School of Software, Dalian Neusoft University of Information, Dalian 116023, China

converting text into oral output in a fix style, but ignore the mixed emotion information from the speaker. It cannot reflect the real emotion state and expression of the speaker, which makes the listener feel the monotony of communication and leads to the deviation of discourse understanding. How to enhance the expressiveness of speech synthesis, especially to make the synthesized speech simulate the emotional state of the speaker, is the development trend of speech synthesis technology, and also the main research content of emotional speech synthesis.

At present, the popular speech synthesis methods mainly include the following categories:

The first method is a class of autoregressive neural network models. WaveNet [8] is a complete convolution autoregressive sequence model, which can generate highly realistic speech samples. SampleRNN [9] is an alternative architecture for unconditional waveform generation, which uses multi-scale recurrent neural networks to explicitly model the original speech at different time resolutions. WaveRNN [10] is a fast autoregressive model based on single layer recurrent neural network. Because these autoregressive models must generate speech in turn, the generation speeds of these models are slower and less inefficient.

The second method is a class of non-autoregressive neural network models, and the most popular method is called generative adversarial networks (GANs) [11]. Because GANs have good learning ability and the ability to simulate data distribution, they have been widely concerned in the field of machine learning. They show excellent performance in image generation [12], image translation [13], image enhancement [14], speech generation [15] and speech conversion [16]. GANs are composed of two neural networks: generator and discriminator. They compete each other through "min–max game". Given a random sample with known prior knowledge, Generator (G) is responsible for generating a pseudo-sample or composite sample $GX \rightarrow Y$. Discriminator (D) is used to distinguish the difference between the generated sample $GX \rightarrow Y$ and the real sample X. The goal of GANs is to train the generator network $GX \rightarrow Y$ which can simulate the real data, so that the discriminator cannot distinguish the real data from the synthetic samples.

The GANs have been successfully applied in the field of speech emotion. Bao et al. [17] designed cycle-consistent adversarial networks (CycleGAN), which generated composite features by bi-directional transformation of emotion feature vector and used the composite corpus to supplement the emotion corpus. In order to make up for the negative effect of the sensitivity of CycleGAN model to spectral features, members of CycleGAN family were born one by one, including CycleGAN VC2 [18], CycleGAN VC3 [19], StarGAN VC [20], MelGAN [21] and MaskCycleGAN [22]. CycleGAN VC2 uses 2–1-2D CNN structure,

which can retain most of the original structure, but it is not suitable for mel-cepstrum conversion. CycleGAN VC3 is an updated version of CycleGAN VC2. It adds time–frequency adaptive normalization (TFAN) structure. Although it improves the performance, it increases the number of converter parameters. MelGAN is the first model that can produce higher-quality speech without additional distillation and perceptual loss. The training parameters and time are acceptable. However, when calculating the loss, the reference terms of MelGAN model are fewer, which often leads to the lack of emotion.

MaskCycleGAN is an extension of MelGAN, which allows the converter to learn the time–frequency structure in a self-monitoring way without additional modules. MaskCycleGAN has a better voice conversion effect between different genders, but MaskCycleGAN also has a weaker emotional voice conversion effect between the same genders.

Aiming at the existing problems of CycleGAN related model, Tian et al. [23] proposed TFGAN model, which can learn from time and frequency domain adversarial. TFGAN discriminate ground-truth waveform from synthetic one in frequency domain for offering more consistency guarantees. Moreover, TFGAN has a set of time-domain loss that help the generator to capture the waveform directly. TFGAN has nearly same synthesis speed as MelGAN, but the fidelity is improved. Therefore, TFGAN is a method to calculate both time domain and frequency domain.

In addition to the above speech enhancement methods, Mixup has been applied to various visual- and speech-related tasks. Tomashenko et al. [24] used Mixup to regularize the acoustic model based on DNN in automatic speech recognition, which improved the performance of speech recognition. Meng [25] et al. designed 'mixspeed' model, which mix the input of two speech and the loss of the corresponding text to realize the transformation of in the two-way corpus. However, few researchers have applied Mixup to speech emotion recognition to enhance the learning and generation of emotion features.

In short, we found that the existing GANs models have their own advantages, but we have not invented a model that can comprehensively consider the expression of different emotions in speech. These existing models only focus on a single aspect of speech synthesis, such as fluency, naturalness, emotion recognition accuracy, robustness, fidelity and system complexity. Considering the problems to be solved in the existing emotional speech synthesis, our goal is to solve the problems of naturalness, robustness, fidelity and emotion recognition accuracy in the process of emotional speech synthesis. Facing our goals, we design an emotion speech synthesis method based on multi-channel time–frequency generative adversarial networks (MC-TFD GANs) and Mixup.

The innovation of our proposed model includes two aspects: MaskCycleGAN with time and frequency domain based on improved GTLU and emotion speech synthesis method based on MC-TFD GANs and Mixup.

(1) Multi-channel time frequency domain GAN model

In the first channel, the time–frequency domain MaskCycleGAN based on the improved GTLU is designed to accelerate the two-way emotional transformation. In the second channel, the time-domain CycleGAN is used to focus on the saliency region of the source signal and highlight the emotional expression of the speech.

(2) Emotion speech synthesis method based on MC-TFD GANs and Mixup

Affective speech synthesis goes through three stages: MC-TFD GANs, loss calculation based on Mixup, and stacking of effective emotion regions based on Mixup. The synthesis method combines multiple channel models with Mixup technology to enhance the feature expression and generation effect of synthetic speech emotion.

In order to improve the reliability and fluency of speech generation, the improved gated tanh linear units (GTLU), multiple loss calculation and emotion saliency region stacking are integrated that provides a new idea for the expansion of emotional corpus.

The structure of this paper is as follows. Part II mainly introduces the MC-TFD GANs model, including time–frequency domain MaskCycleGAN channel based on improved GTLU and time-domain CycleGAN channel based on saliency region. Also, Part II introduces the fusion of MC-TFD GANs and Mixup, including the loss estimation based on Mixup and the effective emotion region weighting method based on Mixup. In the part III, several groups of comparative experiments are designed to provide the results of comparative model training and testing. Part IV is the discussion of the experiment results. Part V is the summary of this work and the prospect of future work.

2 Emotion Speech Synthesis Method Based on MC-TFD GANs and Mixup

2.1 The Synthesis Scheme for Emotion Speech

In order to enhance the advantage of emotion attributes in the generation of emotion speech, a MC-TFD GANs emotion speech synthesis method is designed. The MaskCycleGAN and CycleGAN models are designed for these channels, respectively, and the differences between channels in

time and frequency domain or time domain are reflected. In order to enhance the effect of feature expression and generation of synthetic speech emotion, a channel fusion scheme is designed, which combines MC-TFD GANs model with Mixup technology.

The flow of this method is shown in Fig. 1.

It includes three stages: MC-TFD GANs, loss calculation based on Mixup, and stacking of effective emotion regions based on Mixup.

Stage 1 in the MC-TFD GANs model, two parallel GAN channels are designed for the extraction and generation of time-domain acoustic signals in time–frequency domain and time-domain acoustic information in saliency region.

Stage 2 using Mixup to optimize the linear combination of the channel losses in the stage 1, the mel-cepstrum of the converted speech is generated.

Stage 3 we divide the emotion components of the original speech into different levels. According to the degree of emotion expression, the generated mel-cepstrum is fine-tuned in different levels, and then converted to synthetic a new speech.

Compared with the traditional CycleGAN model, the method proposed does not need to significantly increase the parameters of the generated model and additional voice training, nor does it need to prepare pre-training models such as language model in advance. The execution complexity of our method does not increase significantly, and the operation efficiency is similar to that of CycleGAN.

2.2 Multi-channel Time–Frequency Domain GAN Model

Multi-channel time–frequency domain GANs (MC-TFD GANs) is the core of speech synthesis model, which involves two parallel GAN channels, which are used for the extraction and generation of time–frequency acoustic signals and time-domain saliency region acoustic signals.

In the MC-TFD GANs model, an improved GTLU model is designed and applied for two channels: time–frequency domain MaskCycleGAN based on improved GTLU and time-domain CycleGAN based on saliency region. In the time–frequency MaskCycleGAN channel, GTLU is used to control the degree of emotion feature transformation. In the time-domain CycleGAN channel, only the saliency region of the speech is processed by using ‘image reconstruction’ and ‘emotion region division strategy.’

In order to extract the acoustic features of different scales, we design generator and discriminator models for each channel. They are based on the combination of different convolution layer, gated layer, upsampling and downsampling units.

Fig. 1 The overall process of emotional speech synthesis

2.2.1 Gated Tanh Linear Units (GTLU)

We proposed a new mechanism called gated tanh linear units (GTLU) which is a combination of GTU (gated tanh units) and GLU (gated linear units) [26]. The goal of GTLU is to control the transformation of emotion features.

GTU and GLU are common gating mechanisms. The traditional GTU expression is: $\tanh(X * W + b) \otimes \delta(X * V + c)$, where the activation function ‘tanh’ magnifies the emotion features, which makes the emotion transformation more prominent. However, in the case of gradient back propagation, it is easily to cause gradient explosion and gradient disappearance. The expression of GLU is $\text{relu}(X * W + b) \otimes \delta(X * V + c)$. There is no upper limit requirement for positive input of ReLU, but for the negative input, it is adjusted to zero to avoid gradient disappearance. Based on the characteristics of the gating mechanisms, we combine GTU with GLU to make the transformation of emotion features enlarged without gradient explosion and gradient disappearance.

As one of the smallest units of the model, the GTLU is consist of two 1D CNNs, and the calculation method is shown in Eq. (1).

$$\begin{aligned} h_r &= \text{relu}(X * W + b) \\ h_t &= \tanh(X * W + b) \\ h &= h_r \times h_t \otimes \delta(X * V + c) \end{aligned} \quad (1)$$

where X is the raw wave, and W, V, B, C are the parameters to be learned. $X * W + b$ represents the 1D CNN operation inside the GTLU. h_r is the result of activation function-ReLU calculation after 1D convolution. Similarly, h_t is the result of 1D convolution and activation function Tanh. h is the final result of GTLU after h_r, h_t and activation function-Sigmoid.

2.2.2 MaskCycleGAN with Time and Frequency Domain Based on Improved GTLU (Channel 1)

2.2.2.1 Preprocessing of Acoustic Signal with Filling in Frame The traditional CycleGAN model uses the WORLD vocoder to realize speech conversion. Before training the model, the speech is converted to spectral envelope first. The spectral envelope corresponding from the speech does not have enough ability to capture the correlation information in time and frequency domain, which leads to the insufficient emotion expression of the synthetic speech. In order to alleviate this situation, we use the MelGAN model to obtain the mel-cepstrum of speech, which has strong time–frequency features. In order to enhance the robustness of these features, we superimpose random filling frames over the original mel-cepstrum to generate a new one with filling in frames. On this basis, we redesign the generator and discriminator model, which is called MaskCycleGAN. The MaskCycleGAN structure of channel 1 is shown in Fig. 2. The specific preprocessing process is shown in Fig. 2 with the dotted line.

Fig. 2 The structure of MaskCycleGAN with time and frequency domain (Channel 1)

The preprocessing of MaskCycleGAN is as follows:

- (1) For the mel-cepstrum ‘x’ of the current speech, a mask ‘m’ with the same size as ‘x’ would be created. M contains only 0 or 1. The value of overlapped area in M is 0, and the value of unselected area in M is 1.
- (2) The random frame ‘m’ is superimposed on ‘x’, as is shown in Eq. (2).

$$\hat{x} = x \cdot m \quad (2)$$

where $\cdot$ is element wise product and the mask size is a hyperparameter. Through this superposition, the information of some frames is erased, and they are artificially created by generator.

2.2.2.2 Generator MaskCycleGAN model consists of a generator and a discriminator. In the traditional GAN model, 1D CNN is usually used as the generator. At this time, a frame-level model is used for temporal expansion,

and only the correlation between features within the frame is captured. For speech tasks, the features in mel-cepstrum have a higher dimension. Using 1D CNN for traditional upsampling and downsampling may cause serious degradation. Although the residual model can reduce this distortion, significant noise can still be observed in the synthetic speech.

Different from 1D CNN, 2D CNN is more suitable for converting acoustic features while maintaining the original structure, because it limits the conversion region to local region, which aggravates the local features of the original speech. Based on this, we combine 1D CNN with 2D CNN network to design a multi-level model. Two-dimensional CNN is used to realize upsampling and downsampling, and local features are emphasized. The residual network is implemented by 1D CNN, and the invalid acoustic signals are automatically filtered by the model. Figure 3 shows the structure of the generator. The generator includes three groups of downsampling, seven groups of residual networks and three groups of upsampling.

The downsampling module consists of a 2D CNN and a GTLU. The convolution kernel size is 5 and the stride is 2. The upsampling module includes a 2D CNN, an upsampling module, an InstanceNorm and a GTLU. The convolution kernel size is also 5, and the stride is adjusted to 1. The residual network consists of a pair of 1D CNN, a pair of InstanceNorm and a GTLU. The size of convolution kernel is 3 and the stride is 1. The structure of residual network is shown in Fig. 4.

2.2.2.3 Discriminator In traditional CycleGAN model, only 2D CNN and full connection layer are used to determine the

Fig. 3 The structure of generator

Fig. 4 The structure of residual network

authenticity of generated speech. If we want to get a wider range of receptive field of the discriminator, we need to provide more effective parameters which produce great difficulties to the training model. Based on this, we use GTLU in the MaskCycleGAN model to increase the local information and expand the local receptive field.

The discriminator model consists of a pair of 2D CNN, a pair of GTLU and four groups of downsampling modules. The convolution size is the same as that in the generator. There are couples of CNN and GTLU at the beginning and end, respectively. The structure of the discriminator is shown in Fig. 5.

2.2.3 Channel 2: Time-Domain CycleGAN Based on Saliency Region

In the MaskCycleGAN from the channel 1, the model takes the original acoustics as the input and generates mel-cepstrum based on the filling in frame, which is used for model training. This method emphasizes the time–frequency information of speech signal, which is a global expression of speech emotion in time and space. In order to highlight the expression of specific emotions, we can also extract the emotion expression regions in the speech, and use these regions to fine-tune the model parameters, which can increase the fluency of emotion expression in the generated speech.

In this section, a time-domain CycleGAN model based on saliency region is proposed. The model consists of two stages. The first stage is used to extract significant time-domain regions. In the second stage, CycleGAN model is

Fig. 5 The structure of discriminator

used to train the time-domain speech synthesis model. The structure of channel 2 model is shown in Fig. 6.

2.2.3.1 Stage 1: Extraction of Saliency Region It can quickly lock the saliency region in the speech. Because the shortest time of complete emotion expression only needs several consecutive effective frames, if we can find these sessions, we can use them to replace the whole speech. The session is called saliency region. The emotion expression ability of the other session is weaker. The process of the first stage is as follows.

(1) Acoustic image reconstruction

Acoustic image reconstruction is used for transforming speech into image representation. The generation rule is shown in Eq. (3).

$$t_{i,j} = p_{i * \sqrt{N} + j}, i \in [0, \sqrt{N}], j \in [0, \sqrt{N}] \quad (3)$$

where N represents the dimension of single speech as the original speech. $p_{i * \sqrt{N} + j}$ represents the $i * \sqrt{N} + j$ -dimensional data in the original speech. $t_{i,j}$ is the matrix representation of the original speech. After normalizing $t_{i,j}$, a pixel matrix $t'_{i,j}$ is generated, which is the image representation of the current speech in the time domain. Figure 7 shows the images generated by angry and sad emotion, respectively. We can observe that there are differences in the expression of these emotions.

(2) Emotion saliency region extraction

The data in image representation is the original acoustic signal of speech, which is characterized by timing. When the original acoustic signal has a continuous mutation, there is also a synchronous fluctuation in speech. So, the

Fig. 6 The structure of time-domain CycleGAN (channel 2)

Fig. 7 The image representation of angry and sad class

image form of the speech shows that the signal is mutation (or flat) in the current region. When the local mutation of the image occurs, it is the beginning or ending stage of the speech emotion fluctuation. When the region of the image is flat, the emotion in the current speech does not fluctuate in a large range, that is, the valence of emotion changes less. Based on this, we can use the opportunities of mutation to judge the interval of large-scale energy change. There are differences in the saliency region's expression of different emotions, so locking the saliency regions helps us to identify a particular emotion quickly.

The method to find the saliency regions of emotion expression is as follows:

- (1) A sliding window with a width of $\sqrt{N}$ and a height of $[\sqrt{N}/2, \sqrt{N}]$ is designed. The window can divide the whole image into $[2 * \sqrt{N}, \sqrt{N}]$ regions.
- (2) This window starts from the original position of the image, scans all range of images in turn, and performs differential operation between regions, as shown in Eq. (4) and Eq. (5):

$$\Delta t_{ij} = t'_{i+\sqrt{N}/2,j} - t'_{i,j}, \quad i \in [0, \sqrt{N}], j \in [0, \sqrt{N}] \quad (4)$$

$$\Delta\Delta t_{ij} = \Delta t_{i+1,j} - \Delta t_{i,j}, \quad i \in [0, \sqrt{N}], j \in [0, \sqrt{N}] \quad (5)$$

where t'_{ij} is the normalized pixel matrix obtained by Eq. (3), Δt_{ij} is the difference matrix, and the saliency region is the effective region in Δt_{ij} . Second-order difference is used to find the start and end positions in the saliency region.

- (3) Constraints are added to distinguish flattening and fluctuating regions. Equation (6) and Eq. (7) are used to

Fig. 8 Saliency regions of angry and sad class

calculate t_{start} and t_{end} respectively, as the specific location of the saliency region.

$$t_{\text{start}} = i_1, \Delta\Delta t_{i1,j} > \frac{1}{N} \sum_{i,j=0}^{\sqrt{N}} \Delta\Delta t_{i,j}, \Delta\Delta t_{i1,j}^2 > \frac{1}{N^2} \sum_{i,j=0}^{\sqrt{N}} \Delta\Delta t_{i,j}^2 \quad (6)$$

$$t_{\text{end}} = i_2, \Delta\Delta t_{i2,j} > \frac{1}{N} \sum_{i,j=0}^{\sqrt{N}} \Delta\Delta t_{i,j}, \Delta\Delta t_{i2,j}^2 > \frac{1}{N^2} \sum_{i,j=0}^{\sqrt{N}} \Delta\Delta t_{i,j}^2 \quad (7)$$

- (4) The image and speech in the $[t_{\text{start}}, t_{\text{end}}]$ session is taken out synchronously to form the new image and speech only with saliency region. Figure 8 shows the saliency region image of angry and sad class, respectively. It can be observed that the saliency region of emotions expression is different.

2.2.3.2 Stage 2: Training of Time-Domain Speech Synthesis Model In stage 2, CycleGAN is used to train the time-domain model of speech synthesis and fine-tune the parameters. The structure of stage 2 is shown in Fig. 5.

Firstly, the image representation of the saliency region is transformed into the original speech, and then input into the MelGAN model to obtain the mel-cepstrum of the new speech. Different from the channel 1, we use the CycleGAN model instead of MaskCycleGAN with filling in frame. The structure of generator and discriminator is the same as channel 1.

2.3 Loss Calculation Based on Mixup

2.3.1 Loss Calculation of MC-TFD GANs Model

MC-TFD GANs consists of a MaskCycleGAN model and a CycleGAN model, which share a common ancestor: CycleGAN. CycleGAN originated in the field of computer vision,

which mainly focuses on the translation between unpaired images. It combines adversarial loss and cycle-consistency loss. The goal of CycleGAN is to learn to map $G_{X \rightarrow Y}$ and transform $x \in X$ into $y \in Y$, but it does not depend on the other useful information. When conversing the emotion of speech, CycleGAN also adds the identity-mapping loss to promote the preservation of effective emotion information. The losses in the CycleGAN are as follows:

2.3.1.1 Adversarial loss In order to make the converted new acoustic features indistinguishable from the original features, the adversarial loss is used here, as shown in Eq. (8):

$$L_{\text{adv}}(G_{X \rightarrow Y}, D_Y) = E_{y \sim P_Y(y)}[\log D_Y(y)] + E_{x \sim P_X(x)}[\log(1 - D_Y(G_{X \rightarrow Y}(x)))] \quad (8)$$

where the discriminator D_Y finds the best decision between the real acoustic features and the transformed features by maximizing the loss. The generator $G_{X \rightarrow Y}$ generates features that can cheat the discriminator D_Y by minimizing the loss.

2.3.1.2 Cycle-consistency Loss In adversarial loss, $G_{X \rightarrow Y}(x)$ is required to obey the target distribution. It cannot guarantee the consistency of input and output acoustic features. In order to ensure this mapping, we use the cycle-consistency loss, as shown in Eq. (9):

$$L_{\text{cyc}}(G_{X \rightarrow Y}, G_{Y \rightarrow X}) = E_{x \sim P_X(x)}[\|G_{Y \rightarrow X}(G_{X \rightarrow Y}(x)) - x\|_1] + E_{y \sim P_Y(y)}[\|G_{X \rightarrow Y}(G_{Y \rightarrow X}(y)) - y\|_1] \quad (9)$$

In Eq. (9), we use the positive and negative correlation of double mapping to improve the stability of the model. Cycle-consistency loss helps generators $G_{X \rightarrow Y}$ and $G_{Y \rightarrow X}$ find the best pair of (x, y) combinations in the form of cycle transformation.

2.3.1.3 Identity-mapping loss In order to preserve the expression of the emotion in the speech, we also design the identity-mapping loss, as shown in Eq. (10):

$$L_{\text{id}}(G_{X \rightarrow Y}, G_{Y \rightarrow X}) = E_{x \sim P_X(x)}[\|G_{Y \rightarrow X}(x) - x\|_1] + E_{y \sim P_Y(y)}[\|G_{X \rightarrow Y}(y) - y\|_1] \quad (10)$$

Therefore, the total loss can be written as a linear combination of the three losses, as shown in Eq. (11):

$$L_{\text{full}} = L_{\text{adv}}(G_{X \rightarrow Y}, D_Y) + L_{\text{adv}}(G_{Y \rightarrow X}, D_X) + \lambda_{\text{cyc}} L_{\text{cyc}}(G_{X \rightarrow Y}, G_{Y \rightarrow X}) + \lambda_{\text{id}} L_{\text{id}}(G_{X \rightarrow Y}, G_{Y \rightarrow X}) \quad (11)$$

where L_{full} is the final loss, λ_{cyc} and λ_{id} are hyperparameters used to control the importance of related losses. Equation (11) is used as the loss of CycleGAN model.

2.3.1.4 Loss Calculation Based on MaskCycleGAN In the MC-TFD GANs, the loss calculation method of MaskCycleGAN in channel 1 is different from that of ordinary CycleGAN. The goal of MaskCycleGAN is to learn mapping $G_{\hat{X} \rightarrow Y}^{\text{mask}}$ and transform $\hat{x} \in \hat{X}$ into $y \in Y$. In the same way, we also need to learn to map $G_{Y \rightarrow \hat{X}}^{\text{mask}}$ and transform $y \in Y$ into $\hat{x} \in \hat{X}$, where $\hat{x}$ is a combination of x and m obtained from Eq. (2). Now, the loss of MaskCycleGAN model is shown in Eq. (12).

$$L_{\text{full}} = L_{\text{adv}2}(G_{\hat{X} \rightarrow Y}^{\text{mask}}, D_Y) + L_{\text{adv}2}(G_{Y \rightarrow \hat{X}}^{\text{mask}}, D_{\hat{X}}) + \lambda_{\text{cyc}} L_{\text{cyc}2}(G_{\hat{X} \rightarrow Y}^{\text{mask}}, G_{Y \rightarrow \hat{X}}^{\text{mask}}) + \lambda_{\text{id}} L_{\text{id}2}(G_{\hat{X} \rightarrow Y}^{\text{mask}}, G_{Y \rightarrow \hat{X}}^{\text{mask}}) \quad (12)$$

2.3.2 Loss Calculation Based on Mixup

As a common data enhancement technology, Mixup [18] trains neural networks on the combination of paired examples and their labels. By regularizing the neural network, it can support the simple linear combination of training samples and improve the robustness of synthesis samples. It constructs the following virtual training samples, as shown in Eq. (13) and Eq. (14):

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j \quad (13)$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j \quad (14)$$

where X_i and Y_i are the i -th input and target of the sample. λ is the mixing coefficient calculated from the beta distribution. Mixed data $(\tilde{x}, \tilde{y})$ is represented by combining a set of data samples (X_i, Y_i) , (X_j, Y_j) . Therefore, Mixup extends the training distribution by linear interpolation between training samples and targets. In the fields of computer vision and speech, Mixup can improve the performance of these state-of-the-art systems.

The current model combines Mixup and GAN to realize feature learning and generation in speech emotion. The design scheme is shown in Fig. 9. After obtaining the losses of two channel CycleGAN networks, Mixup is used to combine these losses linearly. Using the feedback regulation of the model, the parameters of CycleGAN are optimized, and the following loss functions are minimized, as shown in Eq. (15):

Fig. 9 Loss calculation based on Mixup

$$L_{\text{full}} = \min(\lambda_{\text{loss}} L_{\text{full1}} + (1 - \lambda_{\text{loss}}) L_{\text{full2}}) \quad (15)$$

where L_{full1} and L_{full2} represent the final losses of the two channels, respectively, $\lambda_{\text{loss}} \in [0, 1]$, λ_{loss} is a random number with beta distribution. The minimum loss can be achieved by adjusting the parameters during training. Only the generator of the channel 1 get the output, and its output content is mel-cepstrum after emotion transformation.

2.4 Effective Emotion Region Stacking Based on Mixup

In this section, we use the mel-cepstrum obtained in Session 3.2 to decode and generate the converted emotion speech. Since emotion is only expression in one (or several) segments of speech, the degree of expression is not the same. Therefore, the effect of conversion can be dynamically adjusted according to the expression of emotion. The design scheme is as follows:

- (1) The mel-cepstrum is aligned with the reconstructed image generated by the saliency region.
- (2) According to the starting position and number of saliency regions, mel-cepstrum is divided into N regions.
- (3) These regions are sorted according to the average value of image pixels. Region NO.1 has the most obvious characteristics. Region NO. N has the least significant characteristics.
- (4) The mel-cepstrum expression of each region is calculated. The expression of the i -th region is shown in Eq. (16):

$$\text{Mel}_i = \frac{(1 - \lambda_{\text{mel}})(N - i)}{N} \text{Mel}_{\text{new}} + \frac{\lambda_{\text{mel}}(N - i)}{N} \text{Mel}_{\text{org}}, \quad i \in [1, N] \quad (16)$$

where Mel_{new} and Mel_{org} are mel-cepstrum after and before conversion, respectively. λ_{mel} is the hyperparameter to generate mel-cepstrum. By segmenting these regions and decoding them with MelGAN model, a new emotion speech can be generated, which can effectively express the target emotion. Finally, the final emotion speech can be generated by arranging and combining generated speech in the original order.

Compared with the traditional CycleGAN model, this method only adds the parameters in the multi-channel, such as λ_{cyc} , λ_{id} , λ_{loss} , Mask size, λ_{mel} and the other parameters. Therefore, the process of parameter fine-tuning has no significant impact on the model training efficiency. The model limits the range of transformed speech to the emotional transformation in two ways, so the model does not need additional speech training, nor does it need to prepare pre-training models in advance, such as language model. On the premise of ignoring the complexity of parameter fine-tuning, the complexity of the current model can be

expressed as the sum of the complexity of a single channel: $O(\text{MaskCycleGAN}) + O(\text{CycleGAN})$.

As we all know, $O(\text{MaskCycleGAN})$ is smaller than $O(\text{CycleGAN})$. Therefore, the complexity of the current model can be approximately expressed as $O(\text{CycleGAN})$.

3 Experiments

3.1 Dataset

All the experiments are carried out on the 8-GPU NVIDIA 1080TI server, using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) Corpus. As the most popular open corpus at present, IEMOCAP includes discrete tags and continuous tags of ten emotions, such as angry, happy, excitement, sad, frustration, fear, surprise, neural and others. Each session in the dataset consists of a dialogue. A male and a female actor execute the script to realize the impromptu dialogue and performance dialogue through emotional scene prompts.

There are 10,039 samples in this corpus, with a total duration of nearly 12 h. Each sample contains speech, video, motion capture and text information. Among them, the samples with at least two-thirds of the same emotional tags would be used as the available data. We use four types of common emotions in the IEMOCAP corpus: happy, sad, angry and neutral. The total numbers of emotion speeches are angry: 1103, happy: 1636, neutral:1708, sad:1084. Different classes of speech can be transformed into each other through the current model. The content of the synthetic speech remains unchanged, and the emotion expression is exchanged.

3.2 Design of Experimental Conditions and Parameters

The goal of the experiment was to analyze the quality of speech emotion expression after conversion. Here, we used MelGAN model to realize speech conversion. In the training of MC-TFD GANs and Mixup models, we used the Pytorch framework to build the model.

The specific design of relevant parameters was as follows:

In channel 1, for each speech, we extracted an 80-dimensional mel-cepstrum with a window size of 1024. In the pretreatment process, mel-cepstrum were normalized. The Adam optimizer was used to train the network for 5.2 k iterations, and the learning rates of generator and discriminator were set to 0.0002 and 0.0001, respectively. The batch size was set to 1, where each speech was consists of 64 randomly cropped frames.

Table 1 List of important hyperparameters

Hyperparameters	Range	Current
Learning rate of generator	0~1	0.0002
Learning rate of discriminator	0~1	0.0001
Batch size	1~256	1
λ_{cyc}	1~20	10
λ_{id}	1~10	5
λ_{loss}	0~1	0.5
λ_{mel}	0~1	0.5
Mask size	1~256	50
Size of convolution kernel	1~32	5
Stride Size	1~16	1
Neuron numbers in the recurrent layer	1~1024	256
Dropout	0~1	0.5

In channel 2, 128 frames were randomly cropped from each speech instead of using the whole sample directly. Batch size, learning rate of generator and discriminator were consistent with channel 1, and momentum term β_1 was set to 0.5.

The hyperparameters involved in the experiment are shown in Table 1.

The verification of speech generation mainly includes two schemes: subjective evaluation and experimental verification. Mean opinion score (MOS) was used in subjective evaluation. In the experimental verification, the external Geneva Minimalistic Acoustic Parameter Set [27] (eGeMAPS) was extracted for each speech (synthetic speech + original speech). The set was consist of 88 high-level descriptors and 22 low-level descriptors features. It contained parameters and extended versions of spectrum and frequency. We designed the bi-directional three-layer long short-term memory (LSTM) model as the verification model with eGeMAPS features as the input. Tensorflow framework was used to build LSTM network. The output of the model was the classification of emotion expression.

3.3 Design of Experiments

The method proposed is tested from four aspects: subjective evaluation, emotional validity of synthetic speech, parameter sensitivity and model convergence.

3.3.1 Subjective Evaluation

In subjective evaluation, listening test was used to evaluate the quality of converted speech. The traditional CycleGAN-VC [18] model was used as the baseline. In order to measure the naturalness of speech, the mean opinion score (MOS) was used here (5: best, 1: worst), and the target speech was taken as the reference. In order to measure the difference of

Table 2 Evaluation results of MOS

MOS	X	A	B
Angry	4.6	4.5	4.8
Happy	4.4	4.2	4.7
Sad	4.7	4.5	4.8
Neutral	4.1	4.1	4.3
WA	4.4	4.35	4.6
UA	4.45	4.325	4.65

emotion expression, XAB test was selected here, where ‘A’ and ‘B’ were the speech converted by the baseline and the method by us, and ‘X’ was the target emotion speech. We randomly selected ten pairs of sentences from the test set and presented them in two orders (AB and BA) to eliminate the deviation caused by the order. For each pair of sentences, the listener was asked to rate ‘X,’ ‘A’ or ‘B’. Ten subjects were invited to participate in the subjective evaluation test. Table 2 shows the MOS of the different emotion speech.

Table 2 shows that the naturalness of speech generated by this model was higher than that generated by baseline. For the emotion with higher activation, such as angry and happy, the naturalness was higher than that of the original speech, and MOS was improved by 4%, which proved that this model had stronger ability in subjective evaluation.

3.3.2 Experiment on Emotion Validity of Synthetic Speech

In order to avoid the impact of different emotion sample class imbalance, we used weighted accuracy (WA) and unweighted accuracy (UA) as indicators to test different emotion classification models. WA was the standard precision calculated over the entire test set. UA was the mean value of the recognition accuracy of all kinds of emotions. For the imbalanced datasets, UA was a more relevant evaluation criteria. Getting a higher UA for speech emotion recognition was one of the challenges of synthetic speech task.

3.3.2.1 Single Channel Experiment This experiment was used to verify the effectiveness of speech synthesis scheme in channel 1. We used the angry and sad class of IEMOCAP as the source and verified the emotion expression through the speech generated by different models. The original speech was used as baseline, and the LSTM model was used as validation model. Compare the training and test results of the following data, as shown in Table 3.

The model of channel 2 did not synthetic speech alone, as channel 2 was only used as an auxiliary tool for model training and parameter tuning. Therefore, we did not test channel 2 alone.

The results in Table 3 showed that the WA and UA of the speech generated by channel 1 were higher, which proved that the speech generated by channel 1 was better

Table 3 Experimental results of single channel

NO	Train	Test	UA	WA
1	Baseline: original speech	Original speech	0.74	0.73
2	Original speech	Synthetic speech by channel 1	0.68	0.67
3	Original speech	synthetic speech by CycleGAN-VC	0.62	0.61
4	Synthetic speech by CycleGAN-VC	synthetic speech by CycleGAN-VC	0.7	0.69
5	Synthetic speech by channel 1	Synthetic speech by channel 1	0.73	0.72
6	Synthetic speech by channel 1	Original speech	0.68	0.68
7	Original speech + synthetic speech by channel 1	Synthetic speech by channel 1	0.7	0.7
8	Original speech + synthetic speech by channel 1	Original speech	0.71	0.69
9	Original speech + synthetic speech by channel 1	Original speech + synthetic speech by channel 1	0.71	0.71

Table 4 Experimental results of MC-TFD GANs model and Mixup

NO	Train	Test	UA	WA
1	Baseline: original speech	Original speech	0.74	0.73
2	Synthetic speech by WaveNet [28]	Synthetic speech by WaveNet	0.64	0.65
3	Synthetic speech by SEGAN [29]	Synthetic speech by SEGAN	0.66	0.66
4	Synthetic speech by CycleGAN VC2 [18]	Synthetic speech by CycleGAN VC2	0.7	0.69
5	Synthetic speech by CycleGAN VC3 [19]	Synthetic speech by CycleGAN VC3	0.71	0.70
6	Synthetic speech by MaskCycleGAN [22]	Synthetic speech by MaskCycleGAN	0.72	0.72
7	Synthetic speech by ours	Synthetic speech by ours	0.77	0.76
8	Original speech	Synthetic speech by ours	0.73	0.73
9	Synthetic speech by ours	Original speech	0.72	0.72
10	Original speech + synthetic speech by ours	Original speech + synthetic speech by ours	0.75	0.75

than CycleGAN-VC in emotion expression. As CycleGAN-VC was very sensitive to mel-cepstrum, the missing signal would easily lead to the increase of the loss. After mixing the original speech and synthetic speech, the accuracy reduction in emotion recognition was less than 3%, which proved that the speech generated by channel 1 was similar with the original speech. However, the speech generated by channel 1 does not have a beneficial improvement effect on speech emotion expression.

3.3.2.2 Multi-channel experiment This experiment was used to verify the effectiveness of MC-TFD GANs and Mixup model. We use the angry and sad emotions in the IEMOCAP dataset as the source. The original speech was used as the baseline, and the LSTM model was used as the validation model. We compared the training and test results of the following data. The experimental results are shown in Table 4.

By comparing the emotional transformation applications of WaveNet, SEGAN, CycleGAN VC2, CycleGAN VC3 and MaskCycleGAN in Table 4, the speech generated by us performed best and has the best WA and UA. The second one was the joint set of original speech and speech generated by us. Comparing the results in Table 4 and Table 3, it could be

Table 5 Experimental results of parameter sensitivity

NO	Hyperparameters	Value	UA
1	$\lambda_{cyc}, \lambda_{id}$	$\lambda_{cyc}=10, \lambda_{id}=5$ (current)	0.75
		$\lambda_{cyc}=5, \lambda_{id}=5$	0.72
		$\lambda_{cyc}=10, \lambda_{id}=3$	0.74
2	λ_{loss}	$\lambda_{loss}=0.5$ (current)	0.75
		$\lambda_{loss}=0.3$	0.73
		$\lambda_{loss}=0.8$	0.7
3	Mask size (m)	m=50	0.75
		m=30	0.72
		m=80	0.73
4	λ_{mel}	$\lambda_{mel}=0.2$ (current)	0.75
		$\lambda_{mel}=0.5$	0.71
		$\lambda_{mel}=0.8$	0.68

found that the accuracy of speech emotion recognition was improved when MC-TFD GANs and Mixup method were import, and the UA is improved by 2.7%, which proved the effectiveness of the MC-TFD GANs and Mixup models in time–frequency domain. The results showed that the generated speech could not only simulate the specific emotion distribution in the original dataset, but also showed more

prominent in emotion expression, which was conducive to improve the accuracy of speech emotion recognition.

3.3.3 Parameter Sensitivity Experiment

In order to test the effectiveness of the hyperparameters in the model, the effects of the following hyperparameters before and after adjustment were compared: λ_{cyc} , λ_{id} , λ_{loss} , Mask size and λ_{mel} . The design scheme is shown in Table 5.

As shown in Table 5, the adjustments of λ_{cyc} , λ_{id} , λ_{loss} had little effect on the performance of emotion recognition. This was because the variables were related to loss calculation and system performance. Mask size and λ_{mel} had an impact on the system performance because they were related to the basic features of speech synthetic. With the increase in λ_{mel} , the synthetic speech was gradually close to the original speech. So λ_{mel} had affected the accuracy of emotion recognition. Mask size directly affects the robustness of the synthetic speech.

3.3.4 Model Convergence Experiment

In order to test the complexity and efficiency of the current model, we compared the loss and convergence cycle of the following models: our method, MaskCycleGAN and CycleGAN. Figure 10 lists the loss of the above models with epochs.

Figure 10 shows that our method, MaskCycleGAN and CycleGAN, could achieve the local stability and final convergence of loss. Through a large number of experiments, it was found that these models had completed the final convergence of loss in 5.2 k, 5 K and 5.5 k iterations, respectively. The convergence speed of the current model was faster than CycleGAN, but slower than MaskCycleGAN. This was because the current model integrated the two models and

Fig. 10 Loss comparison of three models (From left to right: MaskCycleGAN, our method, CycleGAN)

absorbs the complexity characteristics of the two models. However, it should be noted that these models were trained with batch size 1, which was related to the independent transformation of each sample in speech synthesis. However, the execution efficiency of this training method was lower, and the cycle of speech emotion transformation was longer.

4 Discussion

- (1) Through the subjective evaluation experiment and taking the MOS value as a reference, it could be found that the naturalness of the speech generated by the current model was higher than that generated by the baseline. For the higher emotion activation, the MOS value of the generated speech was relatively high. The experimental results showed that the naturalness of highly activated emotion in synthetic speech was significantly higher than that in original audio, and the proposed generation method had a beneficial effect on emotion expression.
- (2) The test of channel 1 showed that the speech generated by this channel was better than the traditional CycleGAN-VC. This was because CycleGAN-VC was very sensitive to mel-cepstrum, and MaskCycleGAN in time–frequency domain in channel 1 could reduce the sensitivity of the model to mel-cepstrum and improve the stability of synthetic speech. After the mixed training of the generated speech of channel 1 and the original speech, the emotion recognition accuracy only had a loss less than 3%, which proved that the emotion expression of the speech generated by channel 1 was similar to that of the original speech. However, the speech generated by channel 1 rarely had a beneficial effect on the speech emotion expression of the original corpus.
- (3) In the multi-channel experiment, compared with the results of other popular speech synthesis models, it could be found that when the multi-channel time–frequency domain GAN and Mixup models were introduced, the accuracy of speech emotion recognition was improved to varying degrees, and its UA was improved by up to 2.7%. This experiment proved that the speech generated by multi-channel time–frequency domain GAN and Mixup model was expressive. It further showed that the generated speech could simulate the specific emotion distribution in the original speech, and the emotion expression was more prominent, which was conducive to the improvement of the accuracy of speech emotion recognition task.
- (4) In the parameter sensitivity experiment, it was proved that mask size and λ_{mel} had affect the accuracy of emotion recognition, because they were related to the basic

features of generated speech, which affected the accuracy and robustness of speech emotion recognition.

- (5) In the model convergence experiment, our method, MaskCycleGAN and CycleGAN could achieve the local stability of loss and finally achieve loss convergence. The convergence rate of our method lied between MaskCycleGAN and CycleGAN. This was because we take a single speech as a unit, the training cycle was longer and the overall training speed was slower.

5 Conclusion

We propose an affective speech synthesis method which combines MC-TFD GANs and Mixup in time and frequency domain, to solve the problem of low naturalness and precision loss of emotion expression generated by GAN. This method is consist of three stages: MC-TFD GANs, loss calculation based on Mixup, and stacking of effective emotion regions based on Mixup.

For MC-TFD GANs, two parallel channels are designed, time–frequency MaskCycleGAN based on improved GTLU and time-domain CycleGAN based on saliency region, which are used to extract global time–frequency acoustic signal and time-domain acoustic information, respectively. Based on Mixup method, the loss estimation of generation model and discriminant model is realized to improve the robustness of the samples. In the third stage, an emotion region stacking method based on Mixup is designed to process the mel-cepstrum of different emotion regions.

The experiments on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) corpus show that the proposed method improves the naturalness and the effectiveness of speech emotion synthesis, and accuracies of the MOS and UA are improved by 3% and 2.7%, respectively. Through a large number of experiments, the effectiveness of speech generated by multi-channel time–frequency domain GAN and Mixup model is proved. It further shows that the generated speech can not only simulate the specific emotional distribution of the original corpus, but also have more prominent emotional expressiveness, and the nature of emotional speech is higher. The corpus expanded by our method can improve the accuracy of speech emotion recognition task.

In the future research, we will improve the emotion speech synthesis model and seek for a more general generative model to make it suitable for emotion interaction and expression across datasets and languages. In addition, in view of the low training efficiency in the process of emotional speech synthesis, we will further reduce the complexity of the model and improve the efficiency of the training

process, so that the model can be used in practical application more effectively.

Acknowledgements This paper was funded by the Liaoning Province Key Laboratory for the Application Research of Big Data, the Dalian Science and Technology Star Project, Grant Number 2019RQ120, and the Intercollegiate cooperation projects of Liaoning Provincial Department of Education, Grant Number 86896244.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Yao, Z.; Wang, Z.; Liu, W., et al.: Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN - ScienceDirect[J]. *Speech Commun.* **120**, 11–19 (2020)
2. Gayathri, P.; Pr Iya, P.G.; Sravani, L., et al.: Convolutional recurrent neural networks based speech emotion recognition[J]. *J. Comput. Theor. Nanosci.* **17**, 3786–3789 (2020)
3. Schuller, B.W.: Speech emotion recognition two decades in a nutshell, benchmarks, and ongoing trends[J]. *Commun. ACM* **61**(5), 90–99 (2018)
4. Mohammed, A.; Carlos, B.: Domain adversarial for acoustic emotion recognition [J]. *IEEE/ACM Trans Audio, Speech, Language Process* **26**, 1–10 (2018)
5. Lu Y.; Mak M W.: Adversarial data augmentation network for speech emotion recognition[C]. In: 2019 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC). IEEE (2019)
6. Ke X.; Cao B.; Bai J., et al.: Speech emotion recognition based on PCA and CHMM[C]. In: 2019 IEEE 8th Joint international information technology and artificial intelligence conference (ITAIC). IEEE (2019)
7. Luo X. L.; Qin X. P.; Jia N.: Research on speech emotion recognition based on MFCC and Its first-order difference features[J]. *Modern Comput.* **11**, 20–24 (2019)
8. Oord A.; Dieleman S.; Zen H et al.: WaveNet: A generative model for raw audio[J]. (2016)
9. Mehri S.; Kumar K.; Gulrajani I., et al.: SampleRNN: An unconditional end-to-end neural audio generation model[J]. (2016)
10. Paul D.; Pantazis Y.; Stylianou Y.: Speaker conditional WaveRNN: towards universal neural vocoder for unseen speaker and recording conditions[C]// INTERSPEECH 2020. (2020)
11. Sahu S.; Gupta R.; Espy-Wilson C.: On enhancing speech emotion recognition using generative adversarial networks[C]// Interspeech. (2018)
12. Song J.: Binary generative adversarial networks for image retrieval[J]. *Int. J. Comput. Vis.* pp. 1–22 (2020)
13. Gong, M.; Chen, S.; Chen, Q., et al.: Generative adversarial networks in medical image processing[J]. *Curr Pharm Des* **27**, 1856–1868 (2021)

14. Praramadhan A. A.; Saputra G. E.: Cycle generative adversarial networks algorithm with style transfer for image generation[J]. (2021)
15. Fernando, T.; Sridharan, S.; McLaren, M., et al.: Temporarily-aware context modelling using generative adversarial networks for speech activity detection[J]. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **28**, 1159–1169 (2020)
16. Li Y.; He Z.; Zhang Y., et al.: High-quality many-to-many voice conversion using transitive star generative adversarial networks with adaptive instance normalization[J]. *J. Circuits, Syst. Comput.*, **12** (2020)
17. Bao, F.; Neumann, M.; Vu, N. T.: Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition. In: *INTERSPEECH*, 2019, pp. 2828–2832
18. Kaneko T.; Kameoka H.; Tanaka K., et al.: CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion[C]. In: *ICASSP 2019 - 2019 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE (2019)
19. Kaneko, T.; Kameoka H.; Tanaka K., et al.: CycleGAN-VC3: Examining and improving CycleGAN-VCs for mel-spectrogram conversion[J] (2020)
20. Kameoka H.; Kaneko T.; Tanaka K et al.: StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks[J]. *IEEE* (2018)
21. Kumar K.; Kumar R.; Boissiere T. D., et al.: MelGAN: Generative adversarial networks for conditional waveform synthesis[J]. (2019)
22. Kaneko, T.; Kameoka, H.; Tanaka, K., et al.: MaskCycleGAN-VC: Learning non-parallel voice conversion with filling in frames[C] (2021)
23. Tian, Q.; Chen, Y.; Zhang Z., et al.: TFGAN: Time and frequency domain based generative adversarial network for high-fidelity speech synthesis[J] (2020)
24. Tomashenko, N. A.; Khokhlov, Y. Y.; Esteve, Y.: Speaker adaptive training and mixup regularization for neural network acoustic models in automatic speech recognition. In: *Interspeech*, pp. 2414–2418 (2018)
25. Meng, L.; Xu, J.; Tan, X.; Wang, J.; Qin, T.; Xu, B.: MixSpeech: Data augmentation for low-resource automatic speech recognition. In: *ICASSP* (2021)
26. Madasu, A.; Rao, V. A.: Gated convolutional neural networks for domain adaptation[C]. In: *International conference on applications of natural language to information systems*
27. Eyben, F.; Scherer, K.R.; Schuller, B.W., et al.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing[J]. *IEEE Trans. Affect. Comput.* **7**(2), 1–1 (2017)
28. Ghosh, S.; Laksana, E.; Morency L. P., et al.: Representation learning for speech emotion recognition[C]. In: *Interspeech*. (2016)
29. Pascual S.; Bonafonte A.; Serrà, J.: SEGAN: Speech enhancement generative adversarial network[J]. (2017)