Endangered Tujia Language Speech Enhancement Research Based on Improved DCGAN

Chongchong Yu[*], Meng Kang, Yunbing Chen, Mengxiong Li, Tong Dai

College of Computer & Information Engineering
Beijing Technology & Business University, Beijing 100048, China
yucc@btbu.edu.cn {1830401006, 10011316215, 1604010312，1604010403}@st.btbu.edu.cn

Abstract. As an endangered language, Tujia language only rely on oral communication. There must exist noises in the process of collecting Tujia language corpus. This paper studies an end-to-end speech enhancement model based on improved deep convolutional generative adversarial network（DCGAN) to extract nearly pure Tujia language speech in noisy environment. Due to the low resource nature of Tujia language, using Chinese corpus as an extension of the Tujia language can effectively solve the problem of insufficient data. The speech enhancement function of the Tujia language was realized using the end-to-end method that consists of symmetric encoding and decoding. By modifying the loss function and network hierarchy parameters, adding the spectrum normalization and imbalanced learning rate made the model more stable during the training process. The experimental results show that the speech enhancement method proposed in this paper can achieve better noise reduction effect on the Tujia language dataset than traditional speech enhancement algorithm and neural network enhancement algorithms.

Keywords: Tujia language, Speech Enhancement, Deep Convolutional Generative Adversarial Network.

1 Introduction

Tujia language contains rich national culture that is passed on from generation to generation by Tujia in China. However, there is no text recording, which has faced the crisis of extinction. In addition, the scope of use of Tujia language is extremely limited, which are usually located in the mountains and valleys with inconvenient traffic. Finding professional recording studios is hard in this environment. The phenomenon that audio files contain noises during the process of recording Tujia language is difficult to avoid [1]. These noises will submerge the useful speech information and impact the subsequent tasks of Tujia language annotation and speech recognition. Removing the noises from the Tujia language speech is a challenge to ensure improving the accuracy of speech recognition and helping phoneticians complete the recording and preservation of endangered languages.

[*] Corresponding author: Chongchong Yu

There are three main methods of speech noise reduction, namely speech enhancement algorithm, using robust speech feature parameters and noise compensation based on model parameter adaptation. Among them, speech enhancement is an effective method to solve noise pollution. Its purpose is mainly two points. The first is to suppress background noise, improve voice quality, and eliminate people's hearing fatigue, which is subjective measurement. The second is to improve the intelligibility of speech, which is an objective measurement [2]. The traditional speech enhancement algorithms have spectral subtraction which has a small amount of calculation and can easily control speech signal distortion and residual noise, but it is easy to exist musical noises [3]. Adaptive filtering, such as Wiener filtering [4] needs to know some features or statistical characteristics of noise. Subspace decomposition based on time domain can also be used for speech enhancement. For example, Chengli S. et al. proposed a signal subspace speech enhancement method based on joint low-rank sparse matrix decomposition, but it had better effect under the condition of low SNR or white noise [5]. In the 1980s, four-layer fully-connected BP network was used for extracting signals from various stationary and non-stationary noises [6]. On this basis, the method of using deep neural network for speech enhancement has also received extensive attention, which has obvious advantages in processing non-stationary noises compared with traditional methods [7,8,9]. However, the deep neural network models are mostly supervised training and rely on a large amount of annotation data. Goodfellow put forward Generative Adversarial Network (GAN), it is not dependent on any priori assumption [10]. At present, GAN has been successfully applied in image processing [11], language text generation [12], audio generation [13] and other aspects. Speech Enhancement GAN (SEGAN) proposed by Pascual et al. obviously reduced noise, but stability and convergence of the model architecture still need further exploration [14]. Alec Radford et al. proposed Deep Convolutional Generative Adversarial Network (DCGAN) for image processing, which used Convolutional Neural Network (CNN) for stabilizing GAN training [15].

In view of the diversity, randomness and non-stationarity of environmental noises in Tujia speech dataset, the speech enhancement model based on improved DCGAN is proposed in this paper, which can carry out rapid enhancement processing, reduce the step of speech feature extraction and realize end-to-end speech enhancement. Because of the low resource nature of Tujia language and the limited amount of data, using Chinese speech dataset as extended dataset can solve the insufficiency of the Tujia language, it is necessary to simplify the network structure and reduce the network depth. The hinge loss function is used for the loss of the model. Moreover, spectral normalization and imbalanced learning rate are added to the model training process. Finally, the PESQ evaluation index is used for evaluating the performance of the model.

The rest of this paper is organized as follows. Section 2 introduces the knowledge of DCGAN. Section 3 demonstrates the proposed approach in detail. The dataset description and experimental results are presented in Section 4. Section 5 concludes our work finally.

2 DCGAN

GAN is a hot research direction in the field of artificial intelligence currently. Many variants have been derived, DCGAN is one of them. Whether it is GAN or its variant, the basic model architecture consists of two neural networks the Generator(G) and the Discriminator(D). G generates new samples that look real by learning input samples. D receives samples from real dataset and the output of G to distinguish the data source. They improve their performance by adversarial learning until generated samples are indistinguishable from real samples. The generator and discriminator of DCGAN adopt improved CNN, which can greatly stabilize the training of GAN.

Different from general CNN, the structure of G in DCGAN becomes fully convolutional network because fully connected hidden layer is cancelled. In addition to the output layer which uses Tanh activation, other layers use ReLU activation. Batch normalization (BN) is added to each layer, which helps to solve the training problems caused by bad network initialization and the flow of gradients in deep networks. D uses ordinary convolution, and all layers use LeakyReLU activation. Similar to G, each layer adds BN operation, and finally fully connected hidden layer is removed.

The training process of DCGAN is the same as ordinary GAN, namely the trainings of G and D are carried out alternately. It is shown in Fig. 1. when G is trained, D remains fixed. Then G accepts a random vector z (assuming that z is subject to some distribution p) to simulate the generation of sample $G(z)$. The error is calculated according to the output of D. Finally, error back propagation algorithm is used to update the parameters of G. when D is trained, G remains fixed. Then the output generator $G(z)$ is taken as negative samples and the real dataset X is taken as positive samples. These positive samples and negative samples are input to D, the error is calculated according to the output of D and the sample labels. Error back propagation algorithm is used to update the parameters of D finally.

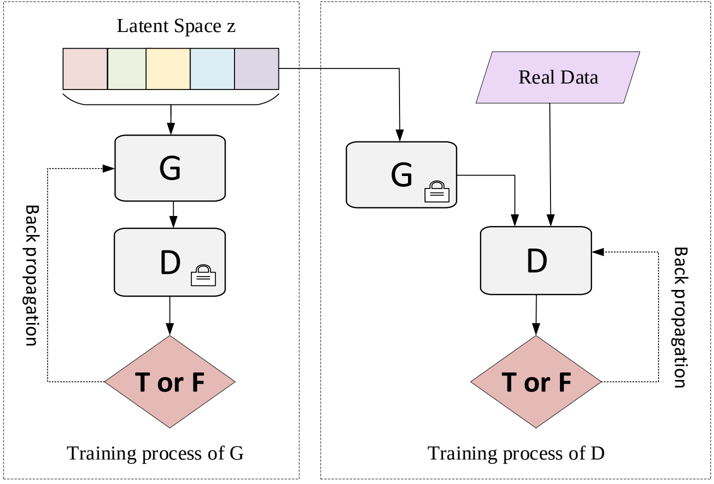

Fig. 1. The training process of GAN

$D(x)$ denotes the probability that D predicts the input data x as real sample. To maximize the expectation $\mathbb{E}_{x\sim p_{data}}$ which the data is from real samples, let D accurately predict $D(x)=1$ when x obeys the probability density of the real samples. To maximize the expectation $\mathbb{E}_{x\sim p_G}$ which the data source is generated samples, let D accurately predicts $D(x)=0$. The purpose of D is to maximize the loss as much as possible. However, the sample distribution generated by G should be as close as possible to real sample distribution, namely $p_G(x)=p_{data}(x)$. D cannot make a judgment on the source of input data so that $D(x)$ is equal to 0.5. The discriminant probability of generated samples $D(G(z))$ needs to be maximized, namely $\log(1-D(G(z)))$ should be minimized. Thus, optimization of the whole network architecture is actually a min-max problem, which is described by the formula as follow:

$$\min_G \max_D V(D,G)=\mathbb{E}_{x\sim \mathrm{p}_{data}(x)}[\log D(x)]+\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \qquad (1)$$

3 the Proposed Approach

3.1 the Speech Enhancement

In this paper, DCGAN is used for study the speech enhancement of Tujia language speech corpus inspired by SEGAN. The main part is G to complete the speech enhancement function. D is responsible for discriminating the generated data from the real data and feeding the result back to the G so that the output of the G is closer to the real data distribution. Until D is difficult to distinguish the authenticity of the input signal, the purpose of removing the noise signal is achieved.

The characteristic of G is the end-to-end structure with encoder-decoder. Encoder consists of convolution layers and PReLU activation. After encoding we can get a vector and connect it to the latent vector z . Decoder consists of deconvolution layers and PReLU activation. After decoding, we can get the enhanced speech waveform. It can be seen that G is designed as a fully convolutional neural network which eliminates the fully connected layer of feature vector classification. This structure allows the network to focus on the time correlation between the input signal and the processing of each layer. In addition, it also reduces the number of training parameters, thus reducing the training time. D is composed of a one-dimensional two-class convolutional network that only outputs the result of judging true and false. The speech enhancement model is shown in Fig. 2.

Skip connection is the response from a convolutional layer is directly propagated to the corresponding mirrored deconvolutional layer [16]. In this way, the information of the waveform can be better transmitted to the decoding stage, so that the reconstructed speech waveform is more refined and the detailed information is not lost after the multi-layer compression.

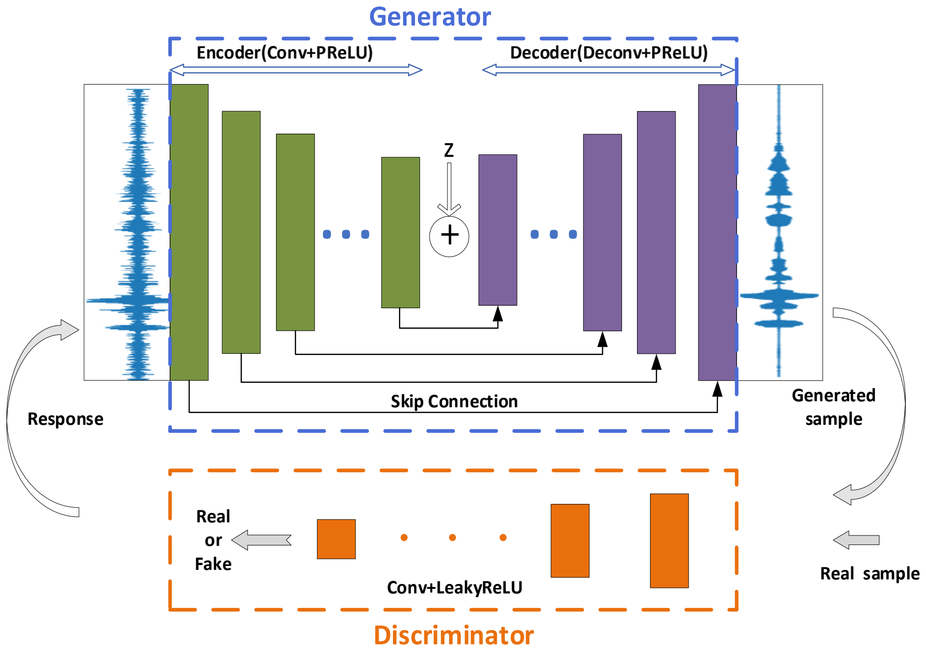

Fig. 2. The speech enhancement model

In order to stabilize the training and improve the quality of the generated samples, the loss function of the model replaces the cross-entropy loss with Hinge loss function [17] which can be better classified. The loss function is defined as follows:

$$L_D = \mathbb{E}_{x \sim p_{data}}[\min(0, -1 + D(x))] + \mathbb{E}_{z \sim p_z}[\min(0, -1 - D(G(z)))] \qquad (2)$$

$$L_G = -\mathbb{E}_{z \sim p_z}[D(G(z))] \qquad (3)$$

Where L_D is the loss function of D and L_G is the loss function of G.

3.2 optimizing train methods

Spectral normalization and imbalanced learning rate were added to the model training process. spectral normalization of generator and discriminator makes it possible to significantly reducing the computational cost of training [18]. The imbalanced learning rate can often be better solved the problem of stability of G and D.

Spectral Normalization. SN performs Singular Value Decomposition (SVD) for the parameter W of every layer of neural network. Every update of W is divided by the largest singular value of W. So, the maximum stretch factor for each layer input x will not exceed 1. It is supposed that after SN every layer of the neural network satisfies:

$$\frac{\|D(x)-D(y)\|}{\|x-y\|} \le K \tag{4}$$

Where $\|\bullet\|$ is L_2 regularization. If K is minimal, then K is called the Lipschitz constraint. However, it is difficult to perform SVD each layer of the neural network in each training iteration, especially when the weight dimension is large. Therefore, power iteration method is adopted to get an approximate solution to the singular value. We initialize a random $\hat{u}$, then update $\hat{u}$ and $\hat{v}$ iteratively according to the follows:

$$\hat{v} \leftarrow \frac{W^T \hat{u}}{\|W^T \hat{u}\|_2} \tag{5}$$

$$\hat{u} \leftarrow \frac{W^T \hat{v}}{\|W^T \hat{v}\|_2} \tag{6}$$

the maximum singular value $\sigma(W)$ of matrix W can be calculated:

$$\sigma(W) \approx \hat{u} W^T \hat{v} \tag{7}$$

Every time the network updates parameters, SN is executed:

$$W \leftarrow \frac{W}{\sigma(W)} \tag{8}$$

Imbalanced learning rate. It is assumed that ω and θ are the parameter vectors of the D and the G. They learn based on the stochastic gradient $\tilde{g}(\theta,\omega)$ of the discriminator loss function L_D and the stochastic gradient $\tilde{h}(\theta,\omega)$ of the generator loss function L_G. There are the actual gradients $g(\theta,\omega)=\nabla_w L_D$ and $h(\theta,\omega)=\nabla_\theta L_G$. The random approximation of the actual gradient is defined according to the random vector $M^{(\omega)}$ and $M^{(\theta)}$:

$$\tilde{g}(\theta,\omega) = g(\theta,\omega) + M^{(\omega)} \tag{9}$$

$$\tilde{h}(\theta,\omega) = h(\theta,\omega) + M^{(\theta)} \tag{10}$$

According to the Two Time Update Rule (TTUR) [19], we use the learning rates $b(n)$ and $a(n)$ for the discriminator and the generator update respectively:

$$\theta_{n+1} = \theta_n + a(n)(h(\theta_n, w_n) + M_n^{(\theta)}) \tag{11}$$

$$w_{n+1} = w_n + b(n)(g(\theta_n, w_n) + M_n^{(w)}) \tag{12}$$

4 Experiments and Result

4.1 Datasets

Due to the limited Tujia language data, we expand the dataset and two datasets are used in the experiment. The first dataset is the Tujia language corpus, which includes 27 oral corpora. There are a total 7830 sentences, with a duration of 7 h, 8 min and 59 s. The Tujia language corpus is divided into two parts, one part contains noises, called the noisy corpus A. The noise types include rooster crowing, chicken crowing, motor vehicle sounds, electronic equipment noises and other noises. The noise fragments are intercepted manually from noisy corpus A using the Elan tool[1]. The details of all kinds of noises are shown in Table 1. Another part is noise-free corpus called clean corpus B. The second dataset is the thchs30 Chinese corpus [20] recorded by 25 people. There are 13395 sentences in total, the recording time is 30 h, the sampling frequency is 16 kHz, and the sampling size is 16 bits.

Table 1. Type and number of Tujia language noises.

Noise Type	Noise Number
Rooster crowing	64
Chicken crowing	31
Motor vehicle sounds	6
Electronic equipment noises	3
Other noises	33

First, the noise segments are added to the clean corpus B and the thchs30 Chinese corpus by the sox tool[2]. Noise injection method is to randomly select the starting position at the sampling point and inject different noises into the thchs30 Chinese corpus according to the proportion of each type of noises to the total number of noises. The formula is as follows:

$$m_{ij} = \frac{N_i M_j}{\sum_{i=1}^{5} N_i}, \quad i = 1, \cdots, 5 \quad , \quad j = 1, \cdots, 25 \tag{13}$$

Where N_i is the number of noise i , M_j is the number of recordings of the j th person in the thchs30 Chinses corpus, m_{ij} is the number of noises injected into the recording of the j th person. The new corpus is called the noisy corpus thchs30. The noisy corpus B is obtained by injecting into the clean corpus B in the same way.

4.2 Experimental Setup

In this study, the Dell PowerEdge R730 server device is used, in which the processor is Intel(R) Xeon(R) CPU E5-2643 v3 @3.40 GHz, the memory size is 64 G, the GPU is NVIDIA Tesla K40 m × 2, and the memory size is 12 GB × 2. The experimental environment for the deep learning framework installed on the Ubuntu 16.04 system is the GPU version of Tensorflow 0.12.

The experimental scheme of speech enhancement based on improved DCGAN is shown in Fig. 3. Firstly, Improved DCGAN (IDCGAN) model is trained by using noisy thchs30 as input and clean thchs30 as output. Then the IDCGAN model is fine-tuned

[1] https://tla.mpi.nl/tools/tla-tools/elan/

[2] http://sox.sourceforge.net/

by using noisy corpus B as input and clean corpus B as output to get Fine-tuning IDCGAN(FIDCGAN) model. Finally, the FIDCGAN model is tested and evaluated with noisy corpus A.

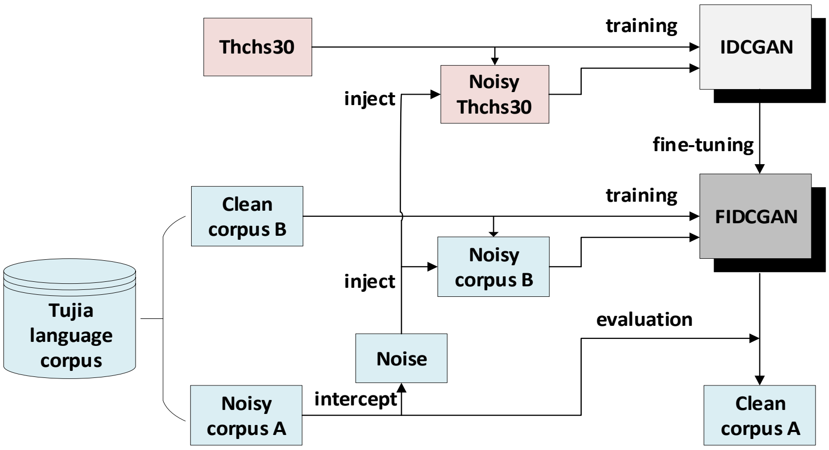

Fig. 3. The experimental scheme of speech enhancement based on improved DCGAN

We extract chunks of waveforms with a sliding window. The window length is about 1 second. The moving distance of the window is 500ms. The encoding stage of G consists of 11 convolutional layers that are one-dimensional strided convolutional layers of filter width 31 and strides of N = 2. The number of filters per layer is 16, 32, 32, 64, 64, 128, 128, 256, 256, 512, 1024. The decoding stage is symmetric with the encoding stage. The D is also one-dimensional convolutional structure. It has two input channels, The LeakyReLU nonlinear activation function with alpha= 0.3 is used, the last layer is the convolution of 1×1, and the output is the result of judging true and false. In IDCGAN model, the learning rate for the D is 0.0003 and the learning rate for the G is 0.0001, use the imbalanced learning rates to train G and D with 1:1 update. The batch size is 24.

Fine-tuning is performed with noisy corpus B and clean corpus B on the basis of IDCGAN model. the learning rate for the D is 0.0002 and the learning rate for the G is 0.00008, the batch size is 16 to obtain the FIDCGAN model. Finally, the FIDCGAN model is tested and evaluated with noisy corpus A.

Except compared with the conventional speech enhancement methods and the speech enhancement methods based on DNN and RNN, Adding BN, SN, BN and SN to each layer of the model in this paper for comparison.

4.3 Experimental Results

Perceptual Evaluation of Speech Quality (PESQ) and Mean Opinion Score of Listening Quality Objective (MOSLQO) are selected as evaluation indexes. PESQ is a typical algorithm in speech quality evaluation. It adopts a linear scoring system with a value between -0.5 and 4.5. The higher the score, the better the quality of speech. PESQ performs level adjustment, input filter filtering, time alignment and compensation, and auditory transformation for he input noisy speech signal and a reference speech signal. Then parameters of the two signals are extracted, and the time-frequency characteristics are integrated to obtain the PESQ score. The MOSLQO value is calculated through the PESQ tool[3]. The evaluation results are shown in Table 2 .

Table 2. The results of different speech enhancement methods.

Method	PESQ	MOSLQO
Spectral Subtraction	1.423	1.309
Wiener Filtering	1.526	1.324
DNN	1.732	1.501
RNN	1.843	1.523
FIDCGAN(BN)	1.921	1.606
FIDCGAN(SN)	2.040	1.664
FIDCGAN(BN+SN)	1.810	1.436

Table 2 shows that the speech enhancement performance proposed in this paper is significant, especially the FIDCGAN model with SN. In addition, the two operations "dividing by the variance" and "multiplying by scaling factor" of BN impact the Lipschitz continuity of the discriminator, the performance will be worse when SN and BN are simultaneously added. Therefore, the BN operation of each layer should be cancelled eventually. The speech spectrum before and after enhancement are shown in Fig. 4.

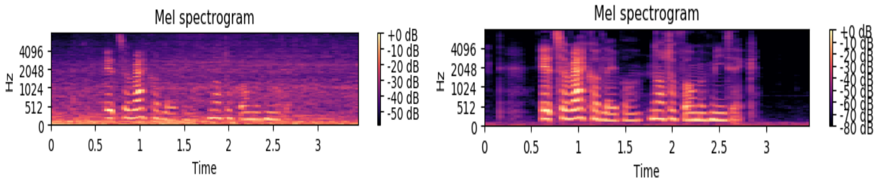

Fig. 4. The speech spectrum before enhancement(left) and after enhancement(right).

The experimental results show that the speech enhancement method based on improved DCGAN can effectively remove the environmental noise in the Tujia language dataset.

[3] https://www.itu.int/rec/T-REC-P.862/en

5 Conclusion

Aiming at the specific scene of the endangered Tujia language, we combine deep CNN and GAN to construct the end-to-end improved DCGAN speech enhancement algorithm. The model preserves phase detail information in the time domain of original speech signal. Spectral normalization and unbalanced learning rates are used to enhance the stability of network training. Compared with the mainstream speech enhancement methods, the experimental results show that the proposed method can obtain better performance. It makes further exploration in the field of speech enhancement and lays a stable foundation for speech feature extraction and recognition. For the low resource Tujia language, we have expanded the data. However, noise type is limited in the specific environment. Therefore, we will attempt to strengthen the generalization of the model and optimize network structure in subsequent work.

Acknowledgment

This research is supported by Ministry of Education Humanities and Social Sciences Research Planning Fund Project, grant number 16YJAZH072, and Major projects of the National Social Science Fund, grant number 14ZDB156.

References

1. Shixuan, X.: On the Recording and Preservation of Endangered Language Data. Journal of Guangxi University for Nationalities (Philosophy and Social Science Edition) 28(5), 11–15 (2006).
2. Hang, H.: Modern Speech signal Processing. Electronic Industry Press, 351-352, Beijing (2014).
3. Dailong, X., Guanyu, L., Ning, M.: Speech Enhancement Research Based on Spectral Subtraction. Journal of Northwest University (Natural Science)38(02), 21-25+87(2017).
4. Navneet, U., Rahul, K.: Single Channel Speech Enhancement: Using Wiener Filtering with Recursive Noise Estimation. Procedia Computer Science, 84(2016).
5. Chengli, S., Jianxiao, X., Yan, L.: A Signal Subspace Speech Enhancement Approach Based on Joint Low-Rank and Sparse Matrix Decomposition. Archives of Acoustics41(2), 245-254(2016).
6. Tamura, S., Waibel, A.: Noise reduction using connectionist models. In: ICASSP 1988, vol.1, pp. 553-556 (1988).
7. Yong, X., Jun, D., Lirong, D., et al.: An Experimental Study on Speech Enhancement Based on Deep Neural Networks. In: *IEEE Signal Processing Letters* 21(1), pp. 65-68(2014).
8. Shi, W., Zhang, X., Sun, M., et al.: Deep neural network based monaural speech enhancement with sparse and low-rank decomposition. In*: IEEE 17th International Conference on Communication Technology (ICCT)*, pp. 1644-1647(2017).
9. Huang, Q., Bao, C., Wang, X., et al.: DNN-Based Speech Enhancement Using MBE Model. In: IWAENC, pp. 196-200 (2018).
10. Goodfellow, I., Pouget-Abadie, M., Mirza, B., et al.: Generative adversarial nets. In Advances in Neural Information Processing Systems (NIPS), pp. 2672–2680 (2014).
11. He, H., Philip S, Y., Changhu, W.: An Introduction to Image Synthesis with Generative Adversarial Nets. In: arXiv: 1803.04469. (2018).

12. Jiaxian, G., Sidi, L., Han, C., et al.: Long Text Generation via Adversarial Training with Leaked Information. In: arXiv: 1709.08624. (2017).
13. Engel, J., Agrawal, K. K., Chen, S., et al.: GANSynth: Adversarial Neural Audio Synthesis. In: ICLR. (2019).
14. Pascual, S., Bonafonte, A., Serra, J.: SEGAN: Speech Enhancement Generative Adversarial Network. In: interspeech. (2017).
15. Alec, R., Luke, M.: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In ICLR. (2016).
16. Xiaojiao, M., Chunhua, S., Yubin, Y.: Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections. In: arXiv: 1603.09056. (2016).
17. Takeru, M., Toshiki, K., Masanori, K., et al.: Spectral Normalization for Generative Adversarial Networks. In: ICLR. (2018).
18. Zhang, H., Goodfellow, I., Metaxas, D., et al.: Self-Attention Generative Adversarial Networks. In: arXiv: 1805.08318. (2018).
19. Heusel, M., Ramsauer, H., Unterthiner, T., et al.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: arXiv: 1706.08500. (2018).
20. Dong, W., Xuewei, Z.: THCHS-30: A Free Chinese Speech Corpus. Computer Science (2015).