SDGAN: Improve Speech Enhancement Quality by Information Filter

Xiaozhou Guo\textsuperscript{1,2}, Yi Liu\textsuperscript{1,2}, Wenyu Mao\textsuperscript{1,2}, Jixing Li\textsuperscript{1,2}, Wenchang Li\textsuperscript{1}, Guoliang Gong\textsuperscript{1,2} and Huaxiang Lu\textsuperscript{1,2,3,4*}

\textsuperscript{1} Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China
\textsuperscript{2} University of Chinese Academy of Sciences, Beijing 100089, China
\textsuperscript{3} CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 200031, China
\textsuperscript{4} Semiconductor Neural Network Intelligent Perception and Computing Technology Beijing Key Lab, Beijing 100083, China
\textsuperscript{*Corresponding author’s e-mail: luhx@semi.ac.cn

Abstract. The speech denoising model based on adversarial generative network has achieved better results than the traditional machine learning model. In this paper, for the short cut connection in the generator, we discuss its influence on the information transfer between encoder and decoder, and propose SDGAN at target. SDGAN sets linear and convolution filters in the short cut connection which adaptively learn the optimal information processing. The information filter still enables the generator to solve the gradient vanishing problem, and it can also avoid information redundancy and improve expression ability. In addition, SDGAN replaces the L1 regularization term in loss function with the L2 regularization term, which not only makes the output speech of the generator closer to the clean speech, but also avoids sparsity. In the experiments, SDGAN significantly performs better than other traditional GAN in five performance metrics (such as PESQ), and the effect of convolution filter is better than that of linear filter.

1. Introduction
In the real environment, the speech signal is usually a mixture of clean speech and background noise signal. In some case, for instance low signal-to-noise ratio (SNR), it is difficult to extract pure speech signal because they have the same frequency bandwidth. And in certain places such as restaurants or malls, noise and pure speech maybe are similar in auditory. Speech enhancement (SE) is a front-end technology to improve the quality and intelligibility of the speech signals, which is employed in automatic speech recognition (ASR), hearing aid and etc.

Previous approaches mainly focus on spectrum characteristics for example spectral subtraction [1]. Power spectrum or magnitude spectrum is usually calculated by the Fourier transform. However these methods ignore the phase characteristics of the speech signal and thus are not satisfactory. Alternatively, there are some models based on statistical learning such as Bayesian estimators [2] and Wiener filters [3] that are also applied in SE task. Analogously, they generally perform well but poorly in the low SNR situation. This is attributed to that the model does not have the ability of modeling the two signals differently. The capacity of expressing and extracting feature has significant influence on the performance, so neural network based model gradually causes concern. This end-to-end model
takes into account amplitude and phase information overall, moreover directly establishes wave-to-wave mapping, which replaces the idea of using the Fourier transform and then operating in the frequency domain. Among them, convolution neural network (CNN) could be utilized to extract deep abstract features [4], recurrent neural network (RNN) can capture the temporal dependence of features, and denoising autoencoder straightly learns the mapping of clean speech and noisy input [5].

generative adversarial network (GAN), a kind of deep generative model, shows promise in multi tasks. The GAN structure consists of two networks including a generator and a discriminator, where the generator is designed to generate fake samples and the discriminator is used to identify the authenticity of the sample. GAN has gained great success in image generation [6], image-to-image translation [7]. For natural language processing, it is applied in text generation [8], machine translation and etc. In the field of speech, GAN also has wide range of utility, such as acoustic model adaptation [9], noise robust [10], voice conversion [11-12], style transfer [13-14]. GAN typically plays a role that maps features from one domain to another (e.g., speaker, styles, clean speech etc.). As to speech enhancement, SEGAN is the first attempt which devises encoder-decoder structure in the generator [15]. In order to stabilize the training, SERGAN substitutes original loss function with the relative loss function [16]. AeGAN proposes CasNet for the generator and a new feature-based loss for the discriminator [17]. TDCGAN introduces a temporal dilated and depthwise separable convolution network and improves speech quality [18]. Adversarial learning is followed in AFM, which combines the acoustic model with the discriminator [19]. However, above approaches only work well under mild noise condition. For one reason, due to the design of network structure and other aspects of the generator, its learning ability is insufficient. For another reason, loss function is unreasonable, which makes the direction of network learning deviate.

In this work, a new generative adversarial network for speech enhancement SDGAN is introduced. Learnable parameters such as linear layer or convolution layer are added in the short cut connection of the generator in proposed SDGAN. The generator would automatically look for the best parameters to control the flow of information from encoder to decoder, which furthermore enhances its learning ability. In addition, a new regularization term with L2 norm is employed to avoid sparse samples, which is closer to the real situation. In other words, above term leads to a more correct direction.

Related experiments about SDGAN are performed on the Edinburgh DataShare dataset. We first train SDGAN using training dataset, we also train SEGAN, SERGAN and WGAN-GP as controlled trials. Then PESQ, LLR, WSS and etc. six evaluation measures are selected to represent the quality of the speech. By comparing the result of speech enhancement finally, SDGAN has advantages in almost every index, which proves its effectiveness.

2. SEGAN

GAN has an excellent performance in generation tasks. The generator \( G \) takes a random noise \( z \) (e.g. Gaussian noise) as input, and generates a fake sample \( G(z) \), which intrinsically learn the mapping from noise distribution to the real sample distribution. Within the discriminator \( D \), both real samples and fake samples are accepted. Then \( D \) outputs a scalar to indicate the degree of confidence, where 1 and 0 represents the real and fake respectively. Through adversarial training, the two networks fall into Nash equilibrium. At this point, the generated sample could fool the discriminator successfully, so that it is difficult to tell true from false.

SEGAN, as depicted in Figure 1, is a speech enhancement technology based on GAN, which operates speech waveform signals on time domain. Different with the typical GANs, the generator of SEGAN replaces Gaussian noise input with noisy speech \( x_{\text{noi}} \), yet gives clean speech \( x_{\text{cle}} = G(x_{\text{noi}}) \). For the generated clean speech \( x_{\text{cle}} \) as well as real clean speech \( \tilde{x}_{\text{cle}} \), the discriminator is responsible to distinguish the difference. If SEGAN reaches satisfied equilibrium, the generator would equip the ability of denoising speech.

SEGAN sets the corresponding loss function which imitates LSGAN. The discriminator selects \( \chi^2 \) divergence to measure the distance between the two distributions, therefore loss function is
Where $\theta_D$ is the weight parameters of $D$. Accordingly, the generator aims at optimizing the learned distance. Besides, additional regularization term is added to make the denoised speech as similar as possible to the pure speech. Hence loss function is

$$\min_{\theta_G} \mathbb{E}_{x_{\text{noi}}} \frac{1}{2} [(D(G(x_{\text{noi}}), x_{\text{noi}}) - 1)^2] + \lambda \| G(x_{\text{noi}}) - x_{\text{cle}} \|_1$$

(2)

Where $\theta_G$ is the weight parameters of $G$, and $\lambda$ is super parameter. These two terms guide the learning process of the generator from different perspectives. The first term is optimized from the level of probability distribution, while the second item is optimized directly from the perspective of sample similarity. Beyond that, JS divergence, Wasserstein distance and etc. can also be used to construct different loss functions.

3. SDGAN

The proposed SDGAN is improved on the basis of SEGAN and has a more sophisticated and reasonable model design. SDGAN will be described next, mainly including the details of the structure and the loss function about the discriminator and generator.

3.1. Generator

The inputs and outputs of the generator are the same as those of SEGAN which is described in section 2. The structure of network is encoder-decoder that is usually effective in generative task as shown in Figure 2. First, the encoder compacts and extracts features of the input noise speech, which is represented as a low dimensional latent vector. Namely, its role here can be understood as eliminating
noise perturbation. Next, the decoder decodes the latent vector and maps it to speech sample. Moreover, short cut connection is added between the corresponding layer in encoder and decoder, increasing the depth of the network and also improving the stability of network training.

The encoder contains a total of 11 layers. Each layer uses a 1-dimensional convolution operation, and the size of its convolution kernel is 31. Since the stride of every convolution is 2, the input speech is compressed by $2^{11}$ times. As the number of layer deepens, the number of convolution kernels increases, which are 16, 32, 32, 64, 64, 128, 128, 256, 256, 512, 1024, so as to improve the richness of features. The pooling operation is removed, but the better performance of the PReLU activation function is selected, whose expression is

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ ax & \text{if } x \leq 0 \end{cases}$$

(3)

Where $a$ is learned parameter that is initialized to value 0.25. Compared with ReLU, PReLU slightly expands the capacity of the network and promotes its representation ability.

The decoder consists of 11 deconvolution layers. Deconvolution is a special convolution operation, which usually aims to expand the dimension of feature graph. Likewise, kernel size and stride of deconvolution are also 31 and 2, which extends the latent vector by $2^{11}$ times. Correspondingly, the number of convolution kernels at each layer is 512, 256, 256, 128, 128, 64, 64, 32, 32, 16, 1, thus end up with a speech sample with the same dimension.

The application of short cut is originally to solve the gradient vanishing problem of deep network. It also has another effect here: when the noisy speech is compressed by the encoder, it will cause a certain amount of information loss, and the lost information may affect the restoration of the speech signal. Therefore, the intermediate features during encoder can be transmitted to the decoder through short cut connection, so that the decoder can collect more information to process. However, if the intermediate feature is transferred directly to the decoder completely, the feature about the noisy part will also be transferred integrally, which is contradictory to the goal of the encoder to eliminate the noise disturbance. We think there is no problem of information loss this moment, but information redundancy. In order to make information flow better, we set linear network or convolution network as information filter in the short cut connection, so that the model can adaptively select the information flow scale according to the training data set. Meanwhile, the size of the information filter needs to be carefully considered. Excessive parameter size affects training efficiency and inference speed, and is prone to over-fitting, while smaller parameter size cannot give full play to the filtering effect. As to linear network filter, it is designed in the form of feature multiplied by weight number, then the value means how much information is retained, and the weights of the same channel are shared. For convolution network filter, it only composed of a convolution network with kernel size of 1, which has stronger screening capability.

Figure 3. structure of the discriminator
3.2. Discriminator
The main network structure of the discriminator is similar to the encoder part of the generator, which consists of 1-dimensional convolution network. Besides, the size of convolution kernel, the number of convolution and the number of convolution kernel are all equal. There are three differences. First, instance regularization technique commonly used in generation model is employed after each convolution, which can not only accelerate the convergence of the model, but also keep the independence between each speech instance. Second, the activation function is set to LeakyReLU which is less expressive and initial value is 0.3. Third, the final output of the discriminator is a scalar value, so a fully connected network is added to the convolution neural network. The specific structure is shown in Figure 3.

3.3. Loss function
Loss function of SDGAN improves the regularization term, which replace 1 norm with 2 norm, that is $\|G(x_{noi}) - x_{cle}\|_2$. Generally speaking, using L1 regularization is easier to obtain sparse solutions, but it is not required that $G(x_{noi})$ and $x_{cle}$ are exactly the same in some dimensions. More importantly, we hope that the two are closer as a whole, so L2 regularization is more in line with the measure requirements. In addition, practice shows that the L2 norm can effectively deter overfitting in training.

4. Experiments
In order to verify the effectiveness of SDGAN, we choose to use Edinburgh DataShare dataset which is adopted in [15]. The training set considers 10 different types of noise, which are added to the clean speech with different SNR (0, 5, 10 and 15 dB). The training set includes 28 speakers, with a total of 11570 utterances. In the test set, 5 different types of noise and 4 different amplitude SNR (2.5, 7.5, 12.5 and 17.5 dB) are combined. The test set covers 2 speakers and 824 utterances. The sampling frequency of any utterance is 16kHz.

In the experiment, we select 5 objective speech evaluation indicators. Distinct with the subjective methods, they use calculation methods to automatically calculate evaluation scores. PESQ (Perceptual Evaluation of Speech Quality) is P.862.2 standard which is recommended by ITU-T. SNR represents the segment signal-to-noise ratio index. IS (Itakura Distance) realized by linear prediction analysis of speech signal which considers gain, while LLR (Log Likelihood Ratio Measure) pays more attention to the similarity of the overall spectral envelope. Additionally, WWS (Weighted Spectral Slope) is also picked.

|                     | PESQ | SNR   | IS     | LLR     | WWS     |
|---------------------|------|-------|--------|---------|---------|
| GAN                 | 2.49 | 17.98 | 9.68   | 0.61    | 32.71   |
| SDGAN (GAN, linear) | 2.55 | 18.10 | 10.32  | 0.63    | 34.12   |
| SDGAN (GAN, convolution) | **2.57** | **18.17** | **10.82** | 0.62   | **38.49** |
| LSGAN               | 2.52 | 17.48 | 9.67   | 0.67    | 32.27   |
| SDGAN (LSGAN, linear) | 2.51 | 17.85 | 10.37  | 0.68    | 32.96   |
| SDGAN (LSGAN, convolution) | **2.55** | **18.11** | **10.59** | 0.64   | **36.87** |
| WGAN                | 2.51 | 18.18 | 9.98   | 0.64    | 30.90   |
| SDGAN(WGAN, linear) | 2.52 | 18.20 | 10.22  | 0.65    | 31.04   |
| SDGAN(WGAN, convolution) | **2.56** | **18.24** | **10.97** | 0.66   | **31.57** |

Related parameters are noted here. The batch size is 128 and epoch is 80. Both the discriminator and generator are trained once respectively within each epoch. We use Adam optimization algorithm in the discriminator and generator with 0.0002 learning rate. We keep the weight parameter of L2 regular term 100 in the loss function of generator. Additionally, in the least square loss function, a, b and c are 1, 0, and 1 respectively. In the Wasserstein distance loss function, the weight of gradient regularization term is 10.
Table 1 shows the result of our SDGAN. No matter what distance metric is used, SDGAN performs clearly better on all five performance metrics, due to the application of information filters in short cut connections whether linear or convolution filter. Besides, compared with linear filter, convolution filter has better speech denoising effect, which indicates that its more powerful information processing ability. Experimental results also confirm that using Wasserstein distance could get more satisfying results, instead of JS or $\chi^2$ divergence.

5. Conclusions
In this work, a speech denoising model named SDGAN based on adversarial generative network is introduced. SDGAN adds an information filter to the short cut connection of the generator, which not only solves the gradient vanishing problem, but also overcomes the redundancy problem of information transfer between the encoder and the decoder. Furthermore, linear and convolution information filters are designed according to the information processing capacity. Moreover, SDGAN replaces L1 regularization term with L2 regularization term to guarantee signal similarity and avoid sparse solution.

To evaluate the validity of the model, we train SDGAN and other GAN models on the Edinburgh DataShare dataset, and select five indicators such as PESQ to measure the effect of speech denoising. SDGAN exhibits a significantly performance comparing to the other GAN based models.

Acknowledgments
This work is supported by National Natural Science Foundation of China U1936106, in part by National Natural Science Foundation of China U19A2080, in part by the CAS Strategic Leading Science and Technology Project (A class) Superconducting Computer Development XDA18040400 High Technology Project 31513070501, in part by Chinese Academy of Sciences STS Project: Development and Application of CMOS Memory and Computer Integrated Neural Network Chip (KFJ-STS-ZDTP-070), in part by Beijing Science and Technology Project: Research on the Core Technology of Brain-like Computing and Cognitive System (Z181100001518006), in part by the Science and Technology Innovation Special Zone Project (1916312ZD00902201). This work is funded by the Beijing Academy of Artificial Intelligence, we would like to express our gratitude.

References
[1] Berouti, M., Schwartz, R., & Makhoul, J. (1979, April). Enhancement of speech corrupted by acoustic noise. In ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 4, pp. 208-211). IEEE.
[2] Loizou, P. C.. (2005). Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum. IEEE Transactions on Speech and Audio Processing, 13(5), 857-869.
[3] Scalart, P. (1996, May). Speech enhancement based on a priori signal to noise estimation. In 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings (Vol. 2, pp. 629-632). IEEE.
[4] Tan, K., & Wang, D. L.. (2018). A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement. INTERSPEECH.
[5] Grzywalski, T., & Drgas, S.. (2019). Using Recurrences in Time and Frequency within U-net Architecture for Speech Enhancement. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.
[6] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., & Ozair, S., et al. (2014). Generative adversarial networks. Advances in Neural Information Processing Systems, 3, 2672-2680.
[7] Harms, J., Lei, Y., Wang, T., Zhang, R., Zhou, J., Tang, X., ... & Yang, X. (2019). Paired cycle - GAN - based image correction for quantitative cone - beam computed tomography. Medical physics, 46(9), 3998-4009.
[8] Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017, February). Seqgan: Sequence generative adversarial nets with policy gradient. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).

[9] Sun, S., Zhang, B., Xie, L., & Zhang, Y.. (2017). An unsupervised deep domain adaptation approach for robust speech recognition. Neurocomputing.

[10] Shinohara, Y.. (2016). Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition. Interspeech 2016.

[11] Kaneko, T., & Kameoka, H.. (2017). Parallel-data-free voice conversion using cycle-consistent adversarial networks.

[12] Hsu, C. C., Hwang, H. T., Wu, Y. C., Tsao, Y., & Wang, H. M.. (2017). Voice Conversion from Unaligned Corpora using Variational Autoencoding Wasserstein Generative Adversarial Networks. Interspeech.

[13] Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L., Roomi, B., Hall, P. (2017) English Conversational Telephone Speech Recognition by Humans and Machines. Proc. Interspeech 2017, 132-136.

[14] Meng, Z., Li, J., Chen, Z., Zhao, Y., Mazalov, V., & Gong, Y., et al. (2018). Speaker-invariant training via adversarial learning.

[15] Pascual, S., Bonafonte, A., & Serrà, Joan. (2017). Segan: speech enhancement generative adversarial network.

[16] Baby, D., & Verhulst, S.. (2019). Sergan: Speech Enhancement Using Relativistic Generative Adversarial Networks with Gradient Penalty. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.

[17] Abdulatif, S., Armansour, K., Guirguis, K., Sajeev, J. T., & Yang, B. (2021, January). Aegan: Time-frequency speech denoising via generative adversarial networks. In 2020 28th European Signal Processing Conference (EUSIPCO) (pp. 451-455). IEEE.

[18] Ye, S., Hu, X., & Xu, X. (2020). Tdcegan: Temporal Dilated Convolutional Generative Adversarial Network for End-to-end Speech Enhancement. arXiv preprint arXiv:2008.07787.

[19] Meng, Z., Li, J., & Gong, Y. (2018). Adversarial feature-mapping for speech enhancement. arXiv preprint arXiv:1809.02251.