Computer Systems Science & Engineering
DOI:10.32604/csse.2021.016524
Article

Tech Science Press

Speech Enhancement via Residual Dense Generative Adversarial Network

Lin Zhou[1,*], Qiuyue Zhong[1], Tianyi Wang[1], Siyuan Lu[1] and Hongmei Hu[2]

[1]School of Information Science and Engineering, Southeast University, Nanjing, 210096, China
[2]Medizinische Physik and Cluster of Excellence "Hearing4all", Department of Medical Physics and Acoustics, University of Oldenburg, 26129, Oldenburg, Germany
[*]Corresponding Author: Lin Zhou. Email: Linzhou@seu.edu.cn

Received: 04 January 2021; Accepted: 17 February 2021

Abstract: Generative adversarial networks (GANs) are paid more attention to dealing with the end-to-end speech enhancement in recent years. Various GAN-based enhancement methods are presented to improve the quality of reconstructed speech. However, the performance of these GAN-based methods is worse than those of masking-based methods. To tackle this problem, we propose speech enhancement method with a residual dense generative adversarial network (RDGAN) contributing to map the log-power spectrum (LPS) of degraded speech to the clean one. In detail, a residual dense block (RDB) architecture is designed to better estimate the LPS of clean speech, which can extract rich local features of LPS through densely connected convolution layers. Meanwhile, sequential RDB connections are incorporated on various scales of LPS. It significantly increases the feature learning flexibility and robustness in the time-frequency domain. Simulations show that the proposed method achieves attractive speech enhancement performance in various acoustic environments. Specifically, in the untrained acoustic test with limited priors, e.g., unmatched signal-to-noise ratio (SNR) and unmatched noise category, RDGAN can still outperform the existing GAN-based methods and masking-based method in the measures of PESQ and other evaluation indexes. It indicates that our method is more generalized in untrained conditions.

Keywords: Generative adversarial networks; neural networks; residual dense block; speech enhancement

1 Introduction

Speech enhancement is widely used in various scenarios, e.g., mobile phones, intelligent vehicles and wearable devices [1]. It is performed as a front-end signal procedure for automatic speech recognition (ASR) [2], speaker identification [3], hearing aid and cochlear implant [4], which aim to improve the intelligibility and quality of target speech in noisy environments [5]. In the early days, speech enhancement was treated as a classical signal processing problem. But now it is formulated as a supervised learning problem from the view of deep learning (DL). Masking-based targets and mapping-based targets are the two groups of training targets for supervised speech enhancement.

The masking-based method focuses on separating clean speech from background interference by measuring speech's binary mask labels in a time-frequency domain. As a result, the mask difference between degraded and clean speeches is adopted as a loss function in the DL training stage. To our knowledge, various mask definitions have been employed in the DL-based speech enhancement, which combine with numerous learning networks. For example, an ideal binary mask (IBM) approach is firstly adopted in the DL-based speech separation [6]. In this work, a pre-trained deep neural network (DNN) is used to estimate the IBM labels on each sub-band. Inspired by IBM, an ideal ratio mask (IRM) method is presented to adaptively fit its mask threshold [7]. Once it is fed to the DNN, speech enhancement performance is significantly improved. Moreover, a long short-term memory (LSTM) network is also tried with IRM. By the temporal dynamics learning of IRM labels, an attractive enhancement result is achieved for binaural speech separation [8]. Besides, the difference on complex ideal ratio mask (cIRM) [9] is another effective loss function for speech enhancement, where cIRM labels provides the real and imaginary components of ideal ratio mask to better confirm the information of clean speech.

The mapping-based method contributes to learning the spectral or temporal representation of clean speech. Existing reports show the mapping-based method has superiority to the masking-based one at a low signal-noise ratio (SNR) [10]. For its training stage, the loss function is directly built by the high-level features learned from degrade and clean speeches. It naturally solves the measure selection problem in the mask-based method. A deep autoencoder (DAE) is firstly proposed to map the Mel-power spectrum between degraded speech and its clean one [11]. Sequentially, the features on log spectral magnitude and log Mel-spectrum are used in DL-based speech separation. [12,13]. As for DNN, it is exploited in the log-power spectrum (LPS) mapping [14]. A similar DNN is carried out to achieve a smaller distortion on various speech levels [15]. A fully convolutional network (FCN) [16] is also introduced to recover clean waveforms of speech. Moreover, convolutional neural networks (CNNs) have already been well adopted into speech signal processing such as binaural sound source localization [17]. Compared with DNN-based methods, the network scale of CNN is greatly reduced.

Among these networks, we notice that generative adversarial networks (GANs) [18–20] are powerful to learn data features with its generative adversarial (GA) loss. As for speech enhancement, an end-to-end GAN model, called as speech enhancement GAN (SEGAN), is suggested [21]. In this model, the input and output data are the waveforms of degraded speech and its estimated clean speech. It shows the waveform-level (temporal) mapping is more concise than the mapping operation in the time-frequency (T-F) domain. Unfortunately, under an unknown (or untrained) noisy condition, the enhancement performance of SEGAN is unsatisfactory. In some cases, it is even worse than that of the masking-based method. To further improve the performance of SEGAN, we present a residual dense GAN (RDGAN). It adopts LPS as the input data instead of the waveform since LPS is perceptually relevant [22,23]. The proposed network consists of two adversarial parts, i.e., generator and discriminator. The main task of generator is to map the degraded LPS to the clean one. To achieve the accurate mapping, we utilize a residual dense block (RDBs) [24] as a component of generator such that it can extract the local features of LPS. Moreover, sequential RDB connections are incorporated on various scales of LPS. It significantly increases the feature learning flexibility and robustness in the time-frequency domain. As for the discriminator, we treat it as a convolutional encoder that outputs the probability of input data similarity. Since the generator tries to fool the discriminator and the discriminator does its best to differentiate, the training procedure of GAN gradually minimizes the GA loss. By combining GAN with T-F representation processed by RDBs, we hope this mapping-based GAN outperforms the masking-based method.

The outline of the paper is organized as follows. In Section 2, the architecture and the implementation of the proposed method are described in detail. Simulation results and analysis are presented in Section 3. Finally, conclusions are drawn in Section 4.

2 Method

In this work, the signal is modeled to be an additive mixture of clean speech signal and noise:

$$y(n) = x(n) + d(n) \tag{1}$$

where $y(n)$, $x(n)$, and $d(n)$ denote noisy speech, clean speech, and noise respectively. n represents the time index.

After framing and windowing, the short-time Fourier transform (STFT) of $y(n)$ can be written as:

$$Y(f,k) = |Y(f,k)|e^{i\theta(f,k)} \tag{2}$$

where $|Y(f, k)|$ and $\theta(f,k)$ are the magnitude and phase of $Y(f, k)$ respectively. The imaginary unit is denoted by i. f and k denote the frequency bin index and the time frame index. With the magnitude, the LPS of the speech signal can be defined as:

$$Y_S(f,k) = 10\log_{10}\left[|Y(f,k)|^2\right] \tag{3}$$

In the proposed method, the generator is designed to map $Y_S(f, k)$ (the noisy LPS) to $\hat{X}_S(f, k)$ (the denoised LPS). Then, to evaluate the quality of the mapping, $X_S(f, k)$ (the clean LPS), $Y_S(f, k)$ and $\hat{X}_S(f, k)$ will be provided to the discriminator, and the output of the discriminator indicates the extent of approximation of $\hat{X}_S(f, k)$, $Y_S(f, k)$ pair and $X_S(f, k)$, $Y_S(f, k)$ pair. During the training, the generator and the discriminator are alternately updated and gradually strengthened because of the gaming between them and the gaming will reach the Nash equilibrium eventually [18]. After that, the well-trained generator can be used to estimate clean LPS from the noisy speech and the obtained mapped LPS together with the original phase can be used to reconstruct the desired enhanced speech.

2.1 *cGAN and LSGAN*

After GAN [18], a batch of derived or modified GANs has been proposed. GANs are generative models learning by adversarial training. Based on GAN, cGAN [19] is a conditioned version of GAN which is offered extra information, in our case, namely the LPS of noisy speech $Y_S(f, k)$. It leads the generator to perform mapping under specified conditions. Furthermore, to solve the vanishing gradient problem caused by sigmoid cross-entropy loss and to guarantee a stable training, LSGAN was proposed [20]. It replaces the cross-entropy loss by the least-squares function. As mentioned in Donahue et al. [25], because the noise variables $p_z(z)$ in the generator tended to be ignored, we decided to discard the latent vector in the generator. Hence, in our case, the objective function can be formulated as Eqs. (4) and (5), where $D(\cdot)$ denotes the probability that the discriminator predicts, and $G(\cdot)$ denotes the output of the generator. E[·] refers to calculating the mean value while $P_{data}(Y_S, X_S)$ means the joint distribution of Y_S and X_S in Eq. (4) and $P_{data}(Y_S)$ means the data distribution of Y_S in Eq. (5).

$$\min_D V_{LSGAN}(D) = \frac{1}{2}E_{Y_S,X_S\sim P_{data}(Y_S,X_S)}\left[(D(X_S,Y_S)-1)^2\right] + \frac{1}{2}E_{Y_S\sim P_{data}(Y_S)}\left[(D(G(Y_S),Y_S))^2\right] \tag{4}$$

$$\min_G V_{LSGAN}(G) = \frac{1}{2}E_{Y_S\sim P_{data}(Y_S)}\left[(D(G(Y_S),Y_S)-1)^2\right] \tag{5}$$

2.2 *Residual Dense Block*

Here, we firstly depict the LPS of clean and noisy speech of consecutive frames in Fig. 1. As we can see, a spectrum can be treated as a feature map with abundant texture. Once noise added, the texture of the speech spectrum will be blurred. Computer vision has achieved great success, employing GANs or other deep learning methods [26,27]. So we tried to make a technology migration. RDB [24] is a network block that can not only obtain the state from the preceding RDB via a contiguous memory (CM) mechanism but also fully utilize all the layers via local dense connections. It contains densely connected layers, local feature fusion (LFF) and local residual learning, together to lead to the CM mechanism. The architecture of one RDB is presented in Fig. 2. Given the input of RDB, F_{input}, the output of the c-th convolutional (Conv) layer of RDB, $F_{c,}$ can be formulated as

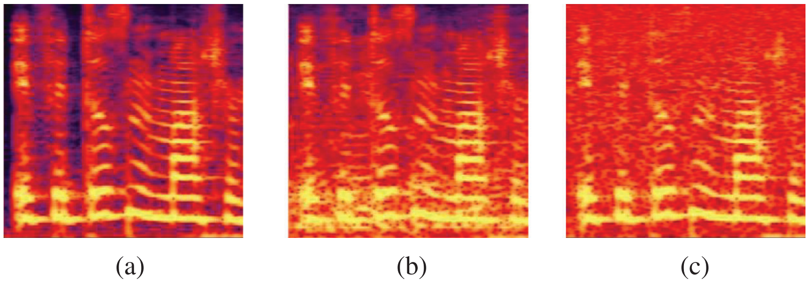

Figure 1: Spectrums of noisy and clean speeches (a) clean speech (b) with babble noise (c) with white noise

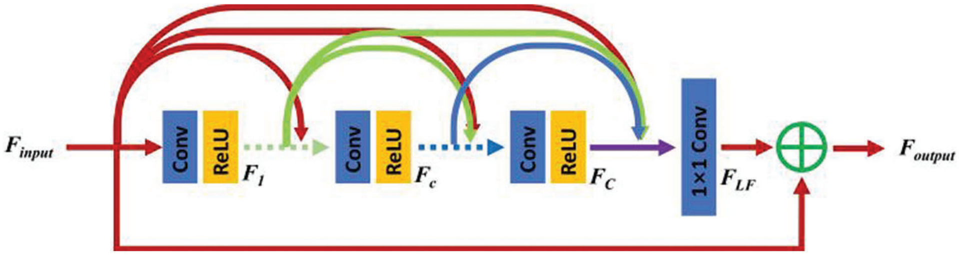

Figure 2: Residual dense block (RDB) architecture

$$F_c = \sigma\left(W_c\left[F_{input}, F_1, \cdots F_{c-1}\right]\right) \tag{6}$$

where σ denotes ReLU [28] activation function and W_c denotes the weights of the c-th Conv layer with bias omitted. Here, all the Conv layers should contain the same number of feature maps. [·] refers to the concatenation of the feature maps.

Then the operation named local feature fusion (LFF) is formulated as

$$F_{LF} = H_{LFF}\left(\left[F_{input}, F_1, \cdots, F_c, \cdots F_C\right]\right) \tag{7}$$

where H_{LFF} denotes the function of the 1×1 Conv layer and the big C denotes the number of the Conv layers. Eventually, the output of RDB can be obtained by

$$F_{output} = F_{input} + F_{LF} \tag{8}$$

2.3 *Proposed Approach: RDGAN*

Fig. 3 shows the generator architecture of the proposed approach. We adopted the well-known U-net [29] structure for the generator, which contains an encoder, a decoder and skip connections. First, the encoder extracts the local and structural features. It halves the size of the feature maps by the convolutional kernel whose stride is 2 instead of by pooling. Each Conv layer is sequentially followed by one ReLU and one instance normalization (IN) layer [30], which significantly reduces the computation load. Also, this process will increase the receptive field of the network, which improves the model robustness. Then the decoder restores the abstract feature and implements dilatation with the Conv layer of which the stride is 1/2. Usually, skip connections play the role of providing the localization information for reconstruction and also allow the combination of local and global feature maps. We integrate the RDBs into skip connections to provide much more details, which exactly suits fine-grained pictures like LPSs. In this way, the encoding and the decoding will not be disturbed and the shortcut of skip connections can remain and provide location information. Hence, the proposed method is named residual dense generative adversarial network (RDGAN).

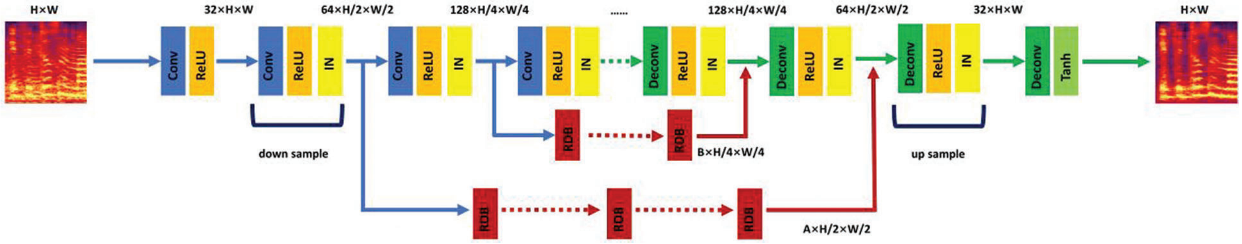

Figure 3: The generator architecture of RDGAN

As for the discriminator, it is similar to the encoding stage of the generator but is a patched version [19,31] to restrict the attention of the structure to local image patches. So, the input of the discriminator will be random patches of the LPS pairs. It must be noted that, instead of ReLU, we use Leaky ReLU [32] following the Conv layers in the discriminator. After the down-sampling of all Conv layers, feature maps will go through a fully connected layer and output a confidence value representing the similarity between input LPS pairs.

Besides the adversarial loss used to handle high-frequencies, the reconstruction loss can also be utilized [24,33,34] to ensure the low-frequency correctness and to minimize the pixel-level difference between the generated and the ground truth. Thus, applying the L1 distance loss, the loss function is finally updated as Eqs. (9) and (10),

$$\min_{D} V_{RDGAN}(D) = \frac{1}{2} E_{Y_S, X_S \sim P_{data}(Y_S, X_S)}\left[(D(X_S, Y_S) - 1)^2\right] + \frac{1}{2} E_{Y_S \sim P_{data}(Y_S)}\left[(D(G(Y_S), Y_S))^2\right] \tag{9}$$

$$\min_{G} V_{RDGAN}(G) = \frac{1}{2} E_{Y_S \sim P_{data}(Y_S)}\left[(D(G(Y_S), Y_S) - 1)^2\right] + \lambda \|G(Y_S) - X_S\|_1 \tag{10}$$

where $\|\cdot\|_1$ denotes calculating the L1 distance and λ is the factor controlling the proportion of the two kinds of loss.

3 Simulation Setup and Result Analysis

3.1 Simulation Setup

To evaluate the proposed method, clean speech signals were taken from the CHAINS *corpus* [35]. The dataset consists of recordings of 36 speakers sharing the same accent obtained in two different sessions with a range of speaking styles and voice modifications. Among the solo reading, four fables told by 9 males and 9 females were used in the training dataset while 33 sentences from the TIMID *corpus* of 3 males and 3 females were used in the test dataset. The speakers of the training and the test dataset are totally different. To build the multi-condition training and test dataset, 4 types of noise (babble, factory, pink, white) from the NOISEX-92 database [36] were added to the mentioned utterances at 4 different SNR, i.e., –5dB, 0dB, 5dB and 10dB. For further tests, 3 untrained types of noise (f16, leopard, Volvo) at SNR –5dB, 0dB, 5dB and 10dB were used to test the robustness of the network. Besides, these noises were also added to the test utterances at untrained SNR level of –7.5dB, –2.5dB, 2.5dB, 7.5dB, 12.5dB. In our simulation, all signals were sampled at 16 kHz.

To obtain the spectrum magnitude, the framing length was 512 with an overlap of 256 samples. After Hamming windowing, 512 points STFT was performed on the frames. Due to the symmetry of the Fourier transform of real discrete sequence, there were 257 frequency bins in the spectrum. After turning magnitude into log-power, we sliced the spectrum every 256 time-frame indexes so the texture could be extracted in a relatively long duration. We also ignored the highest frequency bin to keep the LPS symmetrical. Hence, each LPS was a 256 × 256 image representing the log-power values at T-F units. Based on the RDGAN framework, we utilized 3 down-sample blocks and 3 up-sample blocks in the encoding and decoding stages in the generator. Correspondingly, there were 2 skip connections and we set 6 RDBs in each connection with A as 32 and B as 64 in Fig. 3. Regarding the discriminator, the patch size of the discriminator was set 70 × 70 and 4 down-sample blocks were used with the slope of Leaky ReLU as 0.2. In the generator, except for the first and last Conv layer, the kernel size of whichwas 7 × 7, all the others were 5 × 5. Adam optimizer [37] was used with momentum decay beta 1 and beta 2 set as 0 and 0.9 respectively. λ, the weight of L1, was set to 100. The model was trained for 10 epochs at a batch size of 5.

To evaluate the quality of the enhanced speech, multiple metrics were selected, including the perceptual evaluation of speech quality (PESQ) [38] (from –0.5 to 4.5), mean opinion score (MOS) prediction of the signal distortion attending only to the speech signal (CSIG) (from 1 to 5), MOS prediction of the intrusiveness of background noise (CBAK) (from 1 to 5) and MOS prediction of the overall effect (COVL) (from 1 to 5) [39]. It has been confirmed that the composite measures are more related to subjective evaluation than simple objective measures. The Wiener method based on *a priori* SNR estimation [40] was compared with our method in the matched environments, which means the training dataset and the test dataset have the same SNRs and the same noise types. For unmatched environments in which the test dataset has untrained SNR and untrained noise, the SEGAN [21] model and the IRM-based DNN (IRM-DNN) were compared with the proposed method RDGAN to judge the generalization performance of different methods. We adopted the original setting of SEGAN. For DNN, we used 3 hidden layers of size 2048 to estimate the mask.

3.2 Simulation Results and Analysis

Firstly, we evaluated the performance of the proposed method in the matched noisy environment, that is, the test dataset and the training dataset have the same SNR and same noise type. Tab. 1 shows the comparison of noisy speech, Wiener method and RDGAN method. Here noisy speech means unprocessed speech. It can be seen that RDGAN gets the best scores at all the SNR on all the metrics, which means it not only effectively suppresses distortion and noise but also improves the perceptual quality of enhanced speech. Additionally, the result shows that the proposed RDGAN can improve more under high SNR conditions than under low SNR conditions.

Table 1: Metrics of noisy and enhanced speech in matched environments for different algorithm

Model	Noisy				Wiener				RDGAN			
SNR (dB)	PESQ	CSIG	CBAK	COVL	PESQ	CSIG	CBAK	COVL	PESQ	CSIG	CBAK	COVL
–5	1.03	1.26	1.29	1.09	1.03	1.08	1.30	1.01	**1.07**	**1.33**	**1.63**	**1.13**
0	1.04	1.46	1.53	1.19	1.07	1.30	1.58	1.09	**1.14**	**1.87**	**1.88**	**1.44**
5	1.09	1.80	1.85	1.38	1.16	1.68	1.93	1.31	**1.33**	**2.44**	**2.20**	**1.85**
10	1.20	2.28	2.25	1.71	1.34	2.12	2.29	1.65	**1.61**	**2.93**	**2.53**	**2.25**
average	1.09	1.70	1.73	1.34	1.15	1.54	1.77	1.26	**1.29**	**2.14**	**2.06**	**1.67**

For the unmatched environment, the test dataset has different noise types, compared with the training dataset. Specifically, the noise types in the test dataset are f16, leopard, Volvo noise, while the noise types in the training are babble, factory, pink, white noise. In this simulation, the SNR of the test dataset is the same as that of the training set. As shown in Tab. 2, RDGAN outperforms the Wiener method and SEGAN for all the SNR levels regarding all the metrics used here. But for CBAK, at high SNR, IRM-DNN is slightly superior to RDGAN which indicates IRM-DNN did well in eliminating the intrusiveness of noise. The scores of RDGAN on PESQ and COVL increase more than CSIG and CBAK, which means the perceptual quality of speech is significantly improved. As indicated in Tab. 2, at the high SNR (10dB), the quality of enhanced speech from Wiener, SEGAN and IRM-DNN are lower than that of original noisy speech, which means these models also increased the distortion of the original speech when removing the noise. Although the noise was reduced, the spectral structure of the original speech was also distorted, thereby the perceived quality of the enhanced speech was reduced. This may be because at high SNR the perceptual quality of the original speech is high and the Wiener filtering method is designed for stationary signals, and for non-stationary speech signals, the estimated speech signal correlation function is not proper. The SEGAN method and the IRM-DNN method are also lacking feature extraction. But for RDGAN, with the adversarial loss to handle data distribution, the reconstruction loss to dominate the translation, the quality of enhanced speech is improved even at high SNR.

Tab. 3 presents the results of different methods on untrained noise and untrained SNR. From Tab. 3, the proposed system achieves the best results, which demonstrate the robustness and generalization. Though the CBAK score of IRM-DNN is comparable to RDGAN, RDGAN outperformed IRM-DNN on the whole, meaning the GAN-based mapping method which needs less post-processing can overtake the masking-based method. In addition to the performance improvement, it should be noted that the number of parameters of the RDGAN is about 1/7 of that of SEGAN. As is known to all, a smaller network would help avoid overfitting and assure the generalization of a model. We can say that RDGAN gave the best performance with proper network size and computational advantage compared with SEGAN and IRM-DNN.

Table 2: Metrics of noisy and enhanced speech on unseen noise type

Model	Noisy				Wiener				SEGAN				IRM-DNN				RDGAN			
SNR(dB)	PESQ	CSIG	CBAK	COVL	PESQ	CSIG	CBAK	COVL	PESQ	CSIG	CBAK	COVL	PESQ	CSIG	CBAK	COVL	PESQ	CSIG	CBAK	COVL
–5	1.13	2.37	1.46	1.66	1.21	2.16	1.72	1.59	1.18	1.24	1.66	1.16	1.23	2.31	1.90	1.69	**1.35**	**2.54**	**1.94**	**1.88**
0	1.28	2.73	1.80	1.94	1.36	2.44	2.06	1.81	1.32	1.38	2.00	1.31	1.39	1.62	**2.26**	1.95	**1.53**	**2.93**	2.22	**2.19**
5	1.54	3.19	2.25	2.33	1.56	2.74	2.41	2.08	1.52	1.55	2.37	1.49	1.60	2.90	**2.58**	2.22	**1.79**	**3.33**	2.53	**2.55**
10	1.93	3.68	2.78	2.80	1.80	2.99	2.72	2.35	1.75	1.74	2.71	1.72	1.82	3.12	**2.82**	2.46	**2.09**	**3.69**	2.81	**2.89**
average	1.47	2.99	2.07	2.18	1.48	2.58	2.23	1.96	1.44	1.47	2.18	1.42	1.51	2.74	**2.39**	2.08	**1.69**	**3.12**	2.38	**2.38**

Table 3: Metrics of noisy and enhanced speech on unseen noise type and unseen SNR

Model	Noisy				Wiener				SEGAN				IRM-DNN				RDGAN			
SNR(dB)	PESQ	CSIG	CBAK	COVL	PESQ	CSIG	CBAK	COVL	PESQ	CSIG	CBAK	COVL	PESQ	CSIG	CBAK	COVL	PESQ	CSIG	CBAK	COVL
–7.5	1.08	2.21	1.34	1.56	1.15	2.04	1.58	1.50	1.13	1.19	1.53	1.11	1.17	2.16	1.73	1.57	**1.28**	**2.38**	**1.82**	**1.76**
–2.5	1.19	2.54	1.61	1.79	1.28	2.29	1.89	1.69	1.24	1.30	1.82	1.23	1.30	2.47	**2.08**	1.81	**1.43**	**2.72**	2.07	**2.02**
2.5	1.40	2.95	2.01	2.12	1.46	2.60	2.24	1.94	1.42	1.46	2.18	1.40	1.49	2.77	**2.43**	2.08	**1.65**	**3.13**	2.37	**2.37**
7.5	1.72	3.44	2.50	2.56	1.68	2.87	2.57	2.22	1.63	1.64	2.54	1.61	1.71	3.01	**2.72**	2.34	**1.93**	**3.51**	2.68	**2.72**
12.5	2.16	**3.91**	**3.07**	**3.05**	1.91	3.09	2.86	2.46	1.87	1.83	2.87	1.84	1.91	3.21	2.93	2.56	**2.23**	3.84	2.93	3.05
average	1.51	3.01	2.11	2.21	1.50	2.58	2.23	1.96	1.46	1.48	2.19	1.44	1.52	2.72	**2.38**	2.08	**1.70**	**3.12**	**2.38**	**2.38**

4 Conclusions

We introduce a mapping-based speech enhancement method using GAN and LPS of speech. The proposed RDGAN model integrates residual dense blocks into the encoder-decoder fully-convolutional generator of GAN. It well restores the grain of the two-dimensional T-F representation of speech. The simulation results demonstrate the robustness and generalization of the GAN model and prove the effectiveness of GAN-based mapping approach. However, we seldom consider the noise energy distribution in RDGAN. Thus, our future work will be carried out with more subjective listening evaluations and involve differential processing for different frequency bands, exploring to utilize the phase information and reforming the convolutional networks for efficient mapping of the 2-dimensional T-F representation of speech.

Funding Statement: This work is supported by the National Key Research and Development Program of China under Grant 2020YFC2004003 and Grant 2020YFC2004002, and the National Nature Science Foundation of China (NSFC) under Grant No. 61571106.

Conflicts of Interest: We declare that we have no conflicts of interest to report regarding the present study.

References

[1] J. Park and S. Kim, “Noise cancellation based on voice activity detection using spectral variation for speech recognition in smart home devices,” *Intelligent Automation & Soft Computing*, vol. 26, no. 1, pp. 149–159, 2020.

[2] J. Li, L. Deng, Y. Gong and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[3] A. El-Solh, A. Cuhadar and R. A. Goubran, “Evaluation of speech enhancement techniques for speaker identification in noisy environments,” in *Ninth IEEE Int. Sym. on Multimedia Workshops*, Taichung, Taiwan, pp. 235–239, 2007.

[4] Y. H. Lai, F. Chen, S. S. Wang, X. Lu, Y. Tsao *et al.,* “A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1568–1578, 2017.

[5] P. C. Loizou, *Speech enhancement: theory and practice*. CRC Press, 2007.

[6] Y. Wang and D. Wang, “Boosting classification based speech separation using temporal dynamics,” *13th Annual Conf. of the International Speech Communication Association 2012*, vol. 2, pp. 1526–1529, 2012.

[7] D. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in *Speech Separation by Humans and Machines*, P. Divenyi, Boston, MA: Springer US, pp. 181–197, 2005.

[8] L. Zhou, S. Lu, Q. Zhong, Y. Chen, Y. Tang *et al.,* “Binaural speech separation algorithm based on long and short time memory networks,” *Computers, Materials & Continua*, vol. 63, no. 3, pp. 1373–1386, 2020.

[9] D. S. Williamson, Y. Wang and D. L. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[10] L. Sun, J. Du, L. R. Dai and C. H. Lee, “Multiple-target deep learning for LSTM-RNN based speech enhancement,” in *2017 Hands-Free Speech Communications and Microphone Arrays*, San Francisco, CA, USA: IEEE, pp. 136–140, 2017.

[11] X. Lu, Y. Tsao, S. Matsuda and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Proc. of the Annual Conf. of the Int. Speech Communication Association*, pp. 436–440, 2013.

[12] P. Sen Huang, M. Kim, M. Hasegawa-Johnson and P. Smaragdis, “Joint optimization of masks and deep recurrent neural networks for monaural source separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[13] Y. Xu, J. Du, L. R. Dai and C. H. Lee, “An experimental study on speech enhancement based on deep neural networks,” *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[14] Y. Xu, J. Du, L. R. Dai and C. H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[15] B. Xia and C. Bao, "Wiener filtering based speech enhancement with Weighted Denoising Auto-encoder and noise classification," *Speech Communication*, vol. 60, no. 1, pp. 13–29, 2014.

[16] S. W. Fu, Y. Tsao, X. Lu and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conf.*, Kuala Lumpur: IEEE, vol. 2018-February, pp. 6–12, 2018.

[17] L. Zhou, K. Ma, L. Wang, Y. Chen and Y. Tang, "Binaural sound source localization based on convolutional neural network," *Computers, Materials & Continua*, vol. 60, no. 2, pp. 545–557, 2019.

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley *et al.,* "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 3, no. January, pp. 2672–2680, 2014.

[19] P. Isola, J. Y. Zhu, T. Zhou and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *30th IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2017, pp. 5967–5976, 2017.

[20] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang *et al.,* "Least squares generative adversarial networks," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Venice, Italy: IEEE, vol. 2017-October, pp. 2813–2821, 2017.

[21] S. Pascual, A. Bonafonte and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. of the Annual Conf. of the Int. Speech Communication Association*, vol. 2017-August, pp. 3642–3646, 2017.

[22] S. Tamura, "Analysis of a noise reduction neural network," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Glasgow, UK: IEEE, vol. 3, pp. 2001–2004, 1989.

[23] F. Xie and D. van Compemolle, "A family of MLP based nonlinear spectral estimators for noise reduction," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Adelaide, SA, Australia: IEEE, vol. 2, pp. II53–II56, 1994.

[24] Y. Zhang, Y. Tian, Y. Kong, B. Zhong and Y. Fu, "Residual dense network for image super-resolution," in *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: IEEE, pp. 2472–2481, 2018.

[25] C. Donahue, B. Li and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada: IEEE, vol. 2018-April, pp. 5024–5028, 2018.

[26] K. Fu, J. Peng, H. Zhang, X. Wang and F. Jiang, "Image super-resolution based on generative adversarial networks: a brief review," *Computers, Materials & Continua*, vol. 64, no. 3, pp. 1977–1997, 2020.

[27] W. Chen, T. Sun, F. Bi, T. Sun, C. Tang *et al.,* "Overview of digital image restoration," *Journal of New Media*, vol. 1, no. 1, pp. 35–44, 2019.

[28] X. Glorot, A. Bordes and Y. Bengio, "Deep sparse rectifier neural networks," *Journal of Machine Learning Research*, vol. 15, pp. 315–323, 2011.

[29] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. vol. 9351, pp. 234–241, 2015.

[30] D. Ulyanov, A. Vedaldi and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016. http://arxiv.org/abs/1607.08022.

[31] C. Li and M. Wand, "Precomputed real-time texture synthesis with markovian generative adversarial networks," in *Computer Vision – ECCV 2016*, pp. 702–716, 2016.

[32] A. L. Maas, A. Y. Hannun and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[33] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 2536–2544, 2016.

[34] A. B. L. Larsen, S. K. Sønderby, H. Larochelle and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *33rd Int. Conf. on Machine Learning*, vol. 4, pp. 2341–2349, 2016.

[35] C. Fred, G. Marco, L. Thomas and J. Simko, "The CHAINS corpus: CHAracterizing INdividual Speakers," in *Int. Conf. on Speech and Computer*, pp. 431–435, 2006.

[36] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[37] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” in *3rd Int. Conf. on Learning Representations*, 2015.

[38] ITU-T P.862.2, “Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs,” in *Telecommunication Standardization Sector of Itu*, pp. 12, 2007.

[39] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[40] P. Scalart and J. V. Filho, “Speech enhancement based on a priori signal to noise estimation,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 2, pp. 629–632, 1996.