Qin and Jiang *EURASIP Journal on Wireless Communications and Networking* (2018) 2018:181
https://doi.org/10.1186/s13638-018-1196-0

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

# Improved Wasserstein conditional generative adversarial network speech enhancement

Shan Qin* and Ting Jiang

**Abstract**

The speech enhancement based on the generative adversarial network has achieved excellent results with large quantities of data, but performance in the low-data regime and tasks like unseen data learning still lag behind. In this work, we model Wasserstein Conditional Generative Adversarial Network-Gradient Penalty speech enhancement system and introduce the elastic network into the objective function to simplify and improve the performance of the model in low-resource data environment. We argue that the regularization is significant in learning with small amounts of data and the available information of the input data is key in speech enhancement performance and generalization ability of the model, which means that network parameters and network structure can be set up and designed according to the characteristics of actual input data. Experiments on the noisy speech corpus show that the improved algorithm outperforms previous generative adversarial network speech enhancement approach.

**Keywords:** Wasserstein conditional GAN, Speech enhancement, Generalization

## 1 Introduction

Speech enhancement is one of the main technologies to improve the performance of speech systems in noisy environment [1–3]. Recently, the generation adversarial network has shown great potential in deep learning and has been applied to the field of speech enhancement with large quantities of data, which overcomes the limitations of traditional network for speech enhancement of specific targets and shows good generalization performance for unseen environmental noise [4, 5]. But the problems like instability of learning and mode collapse of generative adversarial network (GAN) affect its practical applications. Many improved algorithms have been proposed such as conditional GAN and Wasserstein GAN to solve those disadvantages. However, those improved GAN algorithms have not been used in speech enhancement yet and the performance in low-resource environment still lags.

Goodfellow et al. introduced the generative adversarial networks (GAN), which formulate a minimax game of a discriminator $D$ and a generator $G$ [6]. The goal is to learn a generator distribution that matches the real data distribution. Martin Arjovsky proposes the Wasserstein GAN to improve the learning stability [7] and Ishaan Gulrajani improves the Wasserstein GAN substituting gradient penalty for weight clipping [8]. Daniel Michelsanti applies conditional GAN for speech enhancement (SE) [9] and Santiago Pascual models SE with GAN [10]. They explore the potential of conditional GAN and GAN for SE and investigate the feasibility and performance of GAN for speech processing.

Compared with most of the current deep learning speech enhancement systems based on the spectrum analysis framework [11–15], GAN and the variant algorithm of GAN work end-to-end with the raw speech data without hand-crafted features extracted and assumptions about the raw data utilized. Because further studies show that outperformance of speech quality is possible, especially when a clean phase spectrum and high frequency information are known [16–18], the raw data retain both phase information and frequency information. GAN and the variant algorithm of GAN provide a sample process without recursive operation in recurrent neural networks and long short-term memory networks [19–22]. But the current GAN system performance in the low-data regime and tasks like unseen data learning still

* Correspondence: tjiang@bupt.edu.cn
Beijing University of Posts and Telecommunications, Haidian District, Beijing 100000, China

lag, because of the instability of training and the problem of gradient disappearing resulting in an inadequate training and the overfitting caused by the complex model.

In this paper, we model WCGAN-GP SE system obtain enhanced speech and introduce an elastic network into the objective function to simplify and improve the performance of the model in low-data regime. Improved WCGAN-GP preserves a variety of features in voice data which is possible to improve the speech quality. And we focus on measuring the ability of models to generalize unseen data and to learn from even very little of it. We explore the factors that influence the enhancement performance of the SE model and the factors related to the generalization of network.

## 2 Wasserstein generative adversarial network

Both $G$ and $D$ are parameterized by convolutional neural networks and the minimax game is given by the following expression:

$$\min \max L(D,G) = E_{x\sim p_{\text{data}(x)}}[\log D(x)] + E_{z\sim p_{g(z)}}[\log(1-D(G(z)))] \quad (1)$$

$p_{\text{data}(x)}$ is the real data distribution and $p_{g(z)}$ is the generated data distribution, $G(z)$ is the generated samples by $G$, $x$ is the real data, and $D(\cdot)$ is the output of discriminator, which is $D(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_g(x)}$. Then, the expansion of Formula 1 is shown:

$$\text{minmax}L(D,G) = E_{x\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(x)}{\frac{1}{2}\left[p_{\text{data}(x)}+p_g(x)\right]}\right] + E_{z\sim p_g}\left[\log \frac{p_g(\text{x})}{\frac{1}{2}\left[p_{\text{data}(x)}+p_g(x)\right]}\right] - 2\log 2 \quad (2)$$

Kullback–Leibler divergence is defined:

$$KL(P_1 \mid |P_2) = E_{x\sim P_1}\log \frac{P_1}{P_2} \quad (3)$$

Jensen-Shannon divergence is

$$JS(P_1 \mid \mid P_2) = \frac{1}{2}KL\left(P_1 \mid \left|\frac{P_1+P_2}{2}\right.\right) + \frac{1}{2}KL\left(P_2 \mid \left|\frac{P_1+P_2}{2}\right.\right) \quad (4)$$

So, the objective function of GAN is as followed:

$$\text{minmax}L(D,G) = 2JS(P_{\text{data}} \mid \left|P_g\right) - 2\log 2 \quad (5)$$

As Jensen-Shannon divergence chose to be a target function, the gradient may be disappearing when the two distributions do not overlap, resulting in the generator unable to learn to improve. This is the main reason for the instability of training.

However, Wasserstein GAN has proved that to minimize reasonable approximation of Earth-Mover distance is the best way to solve the above problems in both theory and practice [23]. The Earth-Mover distance or Wasserstein:

$$W(P_r,P_f) = inf_{\gamma\in\Pi(P_r,P_f)}E_{(x,y)\sim\gamma}[\|x-y\|] \quad (6)$$

$P_r$ and $P_f$ donate the marginal distributions of joint distributions $\gamma(x,y)$. The Earth-Mover distance indicates the cost must be transported from $x$ to $y$ in order to transform the distributions $P_r$ into the distribution $P_f$. The infimum cannot be solved directly, the Lipschitz continuity which is shown in formulate (7) adding a restriction to the continuous function is introduced to the objective function of neural network.

$$|f(x)-f(y) \le K|x-y|| \quad (7)$$

The Lipschitz continuity theorem restricts the maximum local variation of a continuous function. The objection function of Wasserstein GAN is as follows:

$$\min L(D,G)_{|D|_L \le K} = E_{z\sim p_{g(z)}}[D(z)] - E_{x\sim p_{\text{data}(x)}}[D(x)] \quad (8)$$

But weight clipping in the discriminator in Wasserstein GAN results in gradient extinction or gradient explosion, gradient penalty in the discriminator is introduced to make training stability. The gradient penalty is simply added to the Wasserstein distance in the total loss function. To circumvent tractability issues, we enforce a soft version of the constraint with a penalty on the gradient norm for random samples $\hat{x} \sim p_{\hat{x}}$, The objection function of Wasserstein GAN-GP is:

$$\min L(D,G) = E_{z\sim p_{g(z)}}[D(z)] - E_{x\sim p_{\text{data}(x)}}[D(x)] + \lambda E_{\hat{x}\sim p_{\hat{x}}}\left[\left(\|\nabla_z D(\hat{x})\|_2 - 1\right)^2\right] \quad (9)$$

$$\min G_{\text{loss}} = -E_{z\sim p_{g(z)}}[D(z)] \quad (10)$$

where $\hat{x}$ represents a soft version of the constraint with a penalty on the gradient norm for random samples.

$p_{\hat{x}}$ is a uniform sampling distribution along straight lines between pairs of points sampled from the data distribution $p_{\text{data}(x)}$and the generator distribution $p_{g(z)}$.Here, we set $\lambda = 12$, which we found to work well across a variety of architectures and datasets.

However, Wasserstein GAN model learns all the characteristics of the original speech in training data; it is easy to fall into local optimization, too many local features and false characteristics caused by noise, which

results in the generalization and recognition accuracy of the model almost to a valley. In addition, the complex models will result in training overfitting in low-data environment. Accordingly, we propose an improved object function on WCGAN to reduce the complexity of the model and improve the generalization performance, and introduce the Wasserstein conditional GAN SE model in the next section.

## 3 Improved Wasserstein conditional GAN speech enhancement model

The conditional GAN network obtains the desired data for directivity, which is more suitable for the domain of speech enhancement. Therefore, we exploit Wasserstein conditional GAN with GP to implement speech enhancement. There are a positive pair (noisy speech and clean speech) and a negative pair (noisy speech and generate speech), then the discriminator is to distinguish between the two pairs. The structure of speech enhancement conditional GAN is shown in Fig. 1:

Recent several studies have proved that the phase information is significant to improve the effect of the speech intelligibility and quality when spectrograms are used to resynthesized back into time-domain waveforms [24, 25]. As known that the high frequency details of speech also play an important role in speech intelligibility and quality, raw waveform contains the phase features and spectral features of low-frequency and high-frequency of speech. In previous work, Santiago Pascual proposed an end-to-end speech enhancement system based on a generative adversarial network (SEGAN). To research the potential of the WGAN-GP, we continue to explore and improve the speech enhancement system with raw speech data.

The generator is designed to be fully convolution layers without fully connected layers in an auto encoder structure, which may well preserve local information to generate high-frequency components [26]. In the field of ASR, it has been shown that deep learning-based models with raw waveform inputs can achieve higher accuracy than those with hand-crafted features. Therefore, in this paper, the discriminator and the generator are both all full convolutional networks. The fully convolutional layer network structure of the generator is shown in Fig. 2.

Each output sample in FCN relates only locally to the neighboring input regions as shown in Fig. 2. That is distinct from fully connected layers in which the local information and the spatial arrangement of the previous features cannot be well preserved.

The Wasserstein distance insteads of the least mean square error in SEGAN and an elastic network is added to the generator loss to generalize the network performance. The L1 norm performs the automatic selection and sparsity of the features, and the main function of the L2 norm is to avoid overfitting. Elastic network reduces the weight of parameters and simplifies the complexity of models. Ridge regularization keeps more speech feature information to improve intelligibility and

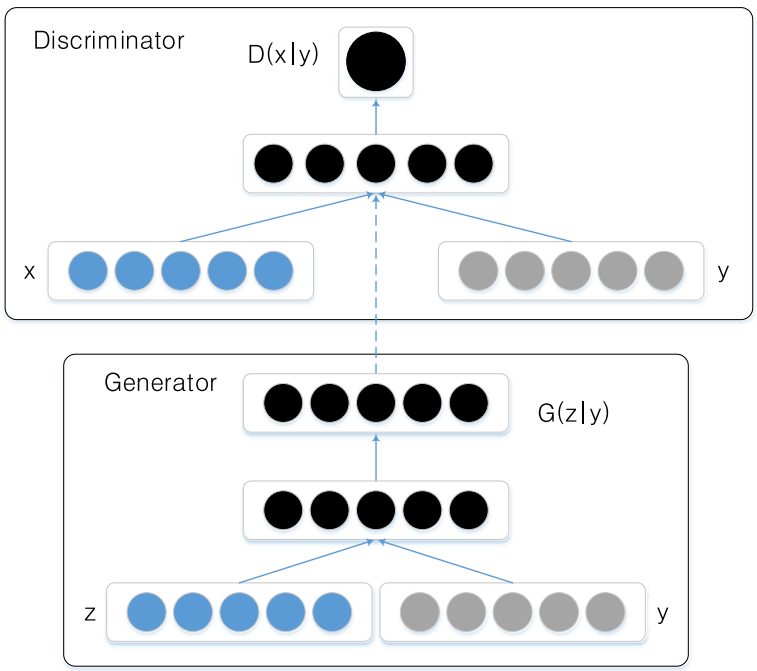


**Fig. 1** The structure of speech enhancement conditional GAN. There are a positive pair (noisy speech and clean speech) and a negative pair (noisy speech and generate speech), then the discriminator is to distinguish between the two pairs

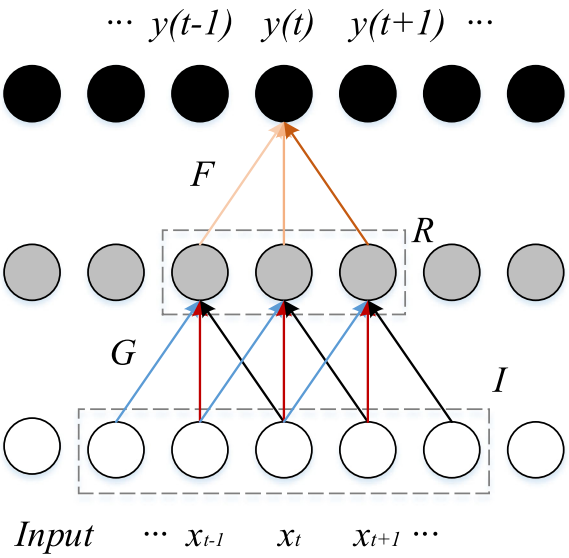


**Fig. 2** Fully convolutional layer network structure. Fully convolutional network preserves local information to generate high-frequency components, and each output sample in FCN relates only locally to the neighboring input regions rather than the total input data

speech quality. The objective function can be represented by the following equations.

$$\min L(D,G) = E_{x_c\sim p_{g(x_c)}}[D(G(x_c),x_c)] - E_{x\sim p_{\text{data}(x)}}[D(x,x_c)] + \lambda E_{x_c\sim p_{g(x_c)}}\left[\left(\|\nabla_z D(G(x_c),x_c)\|_2 - 1\right)^2\right] \tag{11}$$

$$\min G_{\text{loss}} = -E_{x_c\sim p_{g(x_c)}}[D(G(x_c),x_c)] + K[\alpha\|G(x,G(x_c)) - x\|_1 + (1-\alpha)(G(x,G(x_c)) - x)^2] \tag{12}$$

where $x_c$ represents the noisy speech samples. $K$ is the punishment factor. We usually set the $K$ weight of elastic network from 100 to 1000. The constraint condition of L1_ratio is α. In general, it is set according to the experience and experiments.

Lasso and Ridge regularization methods are synthesized in elastic network when many of the features of a data set are interrelated. Elastic network will enhance the group effect between the multiple interrelated variables. The advantage of the tradeoff between Lasso and Ridge is that it allows the stability of Ridge to be inherited during the loop process.

## 4 Experiments and results

### 4.1 Experimental setup

In order to evaluate the performance of WCGAN speech enhancement system, the noisy speech corpus (NOIZEUS) which is developed to facilitate comparison of speech enhancement algorithms among research groups is employed to prepare the training and test sets. The noisy database contains 30 IEEE sentences (produced by three male and three female speakers) corrupted by eight different real-world noises (Babble, Car, Exhibition, Restaurant, Street, Airport, Station, and Train) at different signal-to-noise ratios (SNRs) (0, 5, 10, and 15 dB). There are five sentences for per speaker.

In this study, 1 s of speech with 500 ms overlap was extracted from the waveforms to form a frame for the proposed SEWCGAN model, whose sampling rate is 16 KHz. The proposed SEWCGAN model is trained for 50 epochs with RMS prop and a learning rate of 0.0003, an effective batch size of 100. Due to the addition of the gradient penalty, there is no batch normalization in network structure. In both train and test, a pre-emphasis filter of coefficient 0.95 is applied for all input samples to highlight the high-frequency characteristics to improve the intelligibility and the quality of speech.

Regarding the $\gamma$ weight of gradient penalty, after some experimentation, we set it to 10 for the whole training. We set the $K$ weight of elastic network to 150. Usually, L1_ratio parameters $\alpha(0 < \alpha < 1)$ are employed to regulate the convex combination of L1 and L2. We set it to 0.15 for the whole training. We observed that the $G$ weight follows normal distribution during updating, so the elastic network had practical effect on the learning. As the elastic network value got lower, the quality of the output samples increased, which we hypothesize helped $G$ being more effective in terms of realistic generation.

Regarding the architecture, the structure of improved speech enhancement WCGAN has been shown in Fig. 3. $G$ is composed of 20 convolutional layers of filter width 31 and strides of $N = 2$ which is followed by PReLU activation. The amount of filters per layer increases so that the depth gets larger as the width gets narrower. The resulting dimensions per layer is $16{,}384 \times 1$, $8192 \times 16$, $4096 \times 32$, $2048 \times 32$, $1024 \times 64$, $512 \times 64$, $256 \times 128$, $128 \times 128$, $64 \times 256$, $32 \times 256$, $16 \times 512$, and $8 \times 1024$. It is known that the decoder stage of $G$ is a mirroring of the encoder with the same filter widths and the same amount of filters per layer. The residual network makes training stable and noisy speech vector makes the number of feature maps in every layer to be doubled.

We directly handle the original speech signal and avoid the extraction of acoustic features. Discriminator sent the update information to generator, which fine-tunes the output distribution toward the real distribution to reduce the interference signal.

In this paper, to evaluate the quality of the enhanced speech, we compute the following objective measures. PESQ [− 0.5,4.5]: perceptual evaluation of speech quality. SNR [0, ∞): signal-to-noise ratio. segSSNR [0, ∞): segmental SNR.

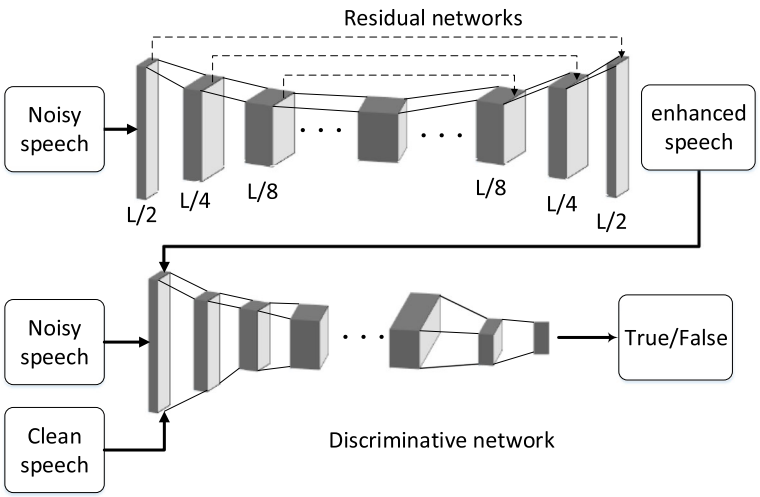


**Fig. 3** The structure of speech enhancement WCGAN. The resulting dimensions per layer is 16,384 × 1, 8192 × 16, 4096 × 32, 2048 × 32, 1024 × 64, 512 × 64, 256 × 128, 128 × 128, 64 × 256, 32 × 256, 16 × 512, and 8 × 1024. The decoder stage of *G* is a mirroring of the encoder with the same filter widths and the same amount of filters per layer. The residual network makes training stable and noisy speech vector makes the number of feature maps in every layer to be doubled

A. Convergent training dataset

To evaluate the convergence of speech enhancement systems in low-resource environment, 900 utterances were randomly selected and corrupted with eight noise types at four SNR levels (0, 5, 10, and 15 dB) as a training set to train the SEGAN model and improved the SEWCGAN model.

B. Generalized training dataset

To explore the influence of the available information on the system generalization performance, we train the improved SE model with different numbers of noise scenes (two scenes, three scenes, four scenes, six scenes, and eight scenes) and different numbers of speakers (two speakers, three speakers, four speakers, and six speakers) at four conditions of SNR. The test data consist of unseen speakers applied by TIMIT database and unseen noisy scenes applied by NoiseX-92.

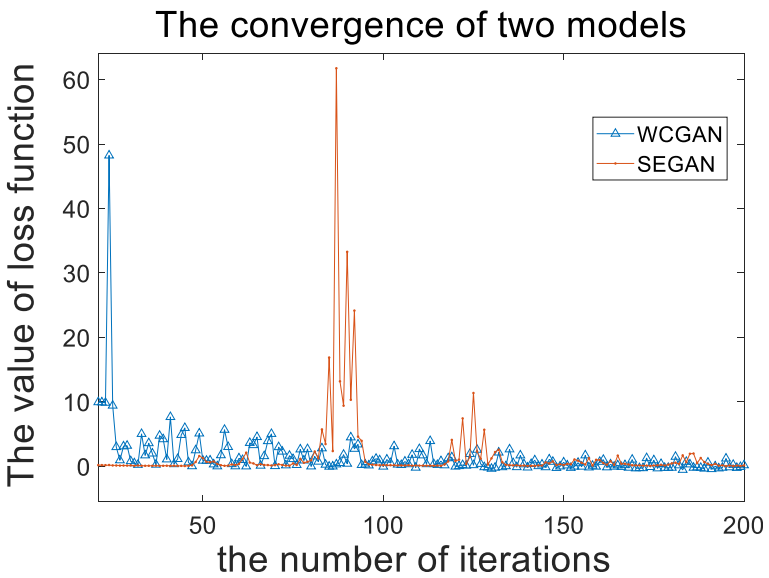


**Fig. 4** The value of discriminator losses. The point labeled line is the loss function value of SEGAN model and the triangular labeled line is the loss function value of improved WCGAN model

## 5 Results and discussion

1. *Comparison of convergence performance*: The convergence of loss function is selected to estimate the convergence of the SEGAN model and improved SEWCGAN model. The experiment results are shown in Fig. 4 as follows.

In Fig. 4, the point labeled line is the loss function value of SEGAN model and the triangular labeled line is the loss function value of improved WCGAN model. As shown in the diagram, the triangular labeled line quickly converges to the stable state near zero, while the convergence rate of the point labeled line is slower. One can observe that the discriminator successfully learns to distinguish the generated representations and the ground-truth in low-resource regime.

Regarding the learning performance of generator, the histogram of update value of the generator is depicted in Fig. 5. Panel (a) is the histogram of the update value of the SEGAN generator, and panel (b) is the histogram of the update value of the improved WCGAN generator. The values in panel (b) mainly range from 0 to 2, and the uneven distribution is easy to cause insufficient learning in network. Relatively, the distribution of the value in panel (a) is uniform which ranges from − 4 to 4, which indicates that the network learning is full and the system performance is better. Compared with SEGAN, the learning performance of the new objective function outperforms the previous one and the convergence rate is faster under the same data conditions.

2. *Comparison and study of generalization performance:* We explore the influence of the available information on the system generalization performance. In this paper, the available information consists of the number of noisy scenes and the number of speakers' clean speech. Firstly, the influence of the number of different training noisy scenes on the system generalization performance is studied. Training data set consists of six speakers' pure voice with different number of noisy scenes (two scenes, three scenes, four scenes, six scenes, and eight scenes) at four different SNR conditions. Test data is applied with a pure male speech in TIMIT database with factor floor noise in NoiseX-92.

Figure 6 depicts the SNR of speech enhancement signal of test data in improved WCGAN models with different number of scenes and SEGAN model with the average SNR. In this part, the main purpose is to explore the impacts of generalization performance, so we show the average SNR of SEGAN only. We find surprisingly

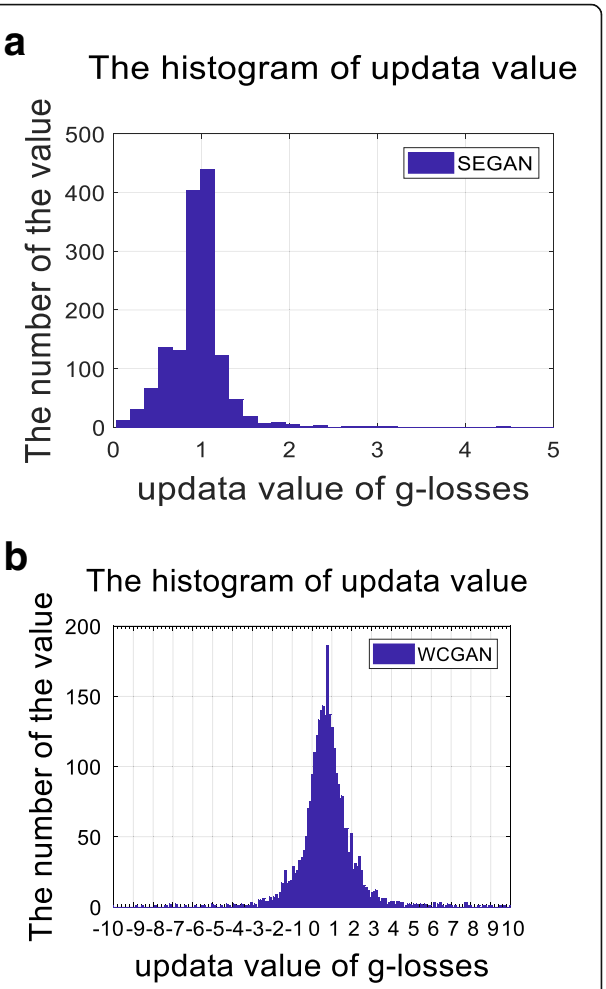


**Fig. 5** The histogram of update value of the loss function for two models. **a** The histogram of the update value of the SEGAN generator. **b** The histogram of the update value of the improved WCGAN generator. The values in **b** mainly range from 0 to 2, and the uneven distribution is easy to cause insufficient learning in network. Relatively, the distribution of the value in **a** is uniform which ranges from − 4 to 4, which indicates that the network learning is full and the system performance is better

that, simply by applying elastic network to the WCGAN models, our improved model achieves mean value 23.65 gain in the SNR on WCGAN and achieves 21.05 gain on SEGAN. The performance of the four-scene model and eight-scene model are almost the same, which are better than other three models. Figure 7 describes the segmented SNR of enhanced speech in different scenes in two models. By calculating the mean, the improved model achieves mean value 4.28 gain in the segSNR on WCGAN and achieves 1.45 gain on SEGAN. The four-scene model outperforms the other models in segmented SNR.

According to the two-graph information, it intimates that the four-scene SE model outperforms the other scene models. It means that if the number of clean speech is certain, increasing the noisy training data does

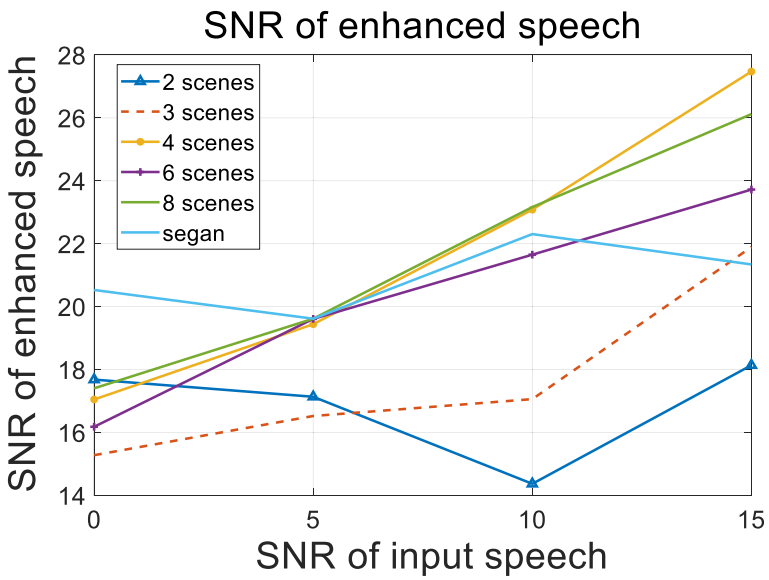


**Fig. 6** The SNR of enhancement speech. It depicts the SNR of speech enhancement signal of test data in improved WCGAN models with different numbers of scenes and SEGAN model with the average SNR. We find surprisingly that the improved model achieves mean value 23.65 gain in the SNR on WCGAN and achieves 21.05 gain on SEGAN

not play a totally positive role in the enhancement performance and generalization performance of the speech system.

Secondly, the influence of the number of speakers' clean speech on the system generalization performance is discussed. Training data set consists of four scenes with different numbers of speakers' clean speech (two speakers, three speakers, four speakers, and six speakers) at four different SNR conditions. Test data is as the same as the above experiment. The experimental results are shown in Figs. 8 and 9 as follows.

Figure 8 shows the SNR of enhanced speech of test data in improved WCGAN models with different numbers of speakers and SEGAN model with the average SNR. The Performance of the four-speaker model and the six-speaker model are almost the same at low SNR condition. The six-speaker model performs prominently at high SNR conditions. The SEGAN model works well at 0 dB condition as well. Figure 9 describes the segmented SNR of enhanced speech in different models. The four-speaker model shows a better performance than other models in low SNR of input

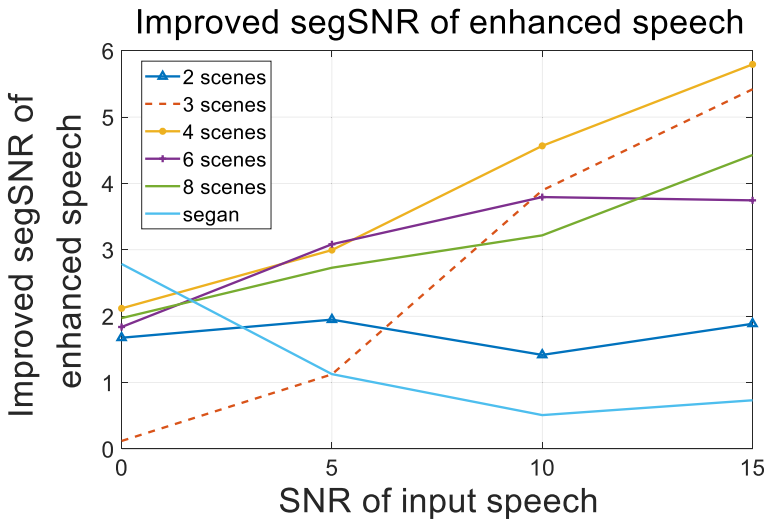


**Fig. 7** The segSNR of enhancement speech. It describes the segmented SNR of enhanced speech in different scenes in two models. By calculating the mean, the improved model achieves mean value 4.28 gain in the segSNR on WCGAN and achieves 1.45 gain on SEGAN

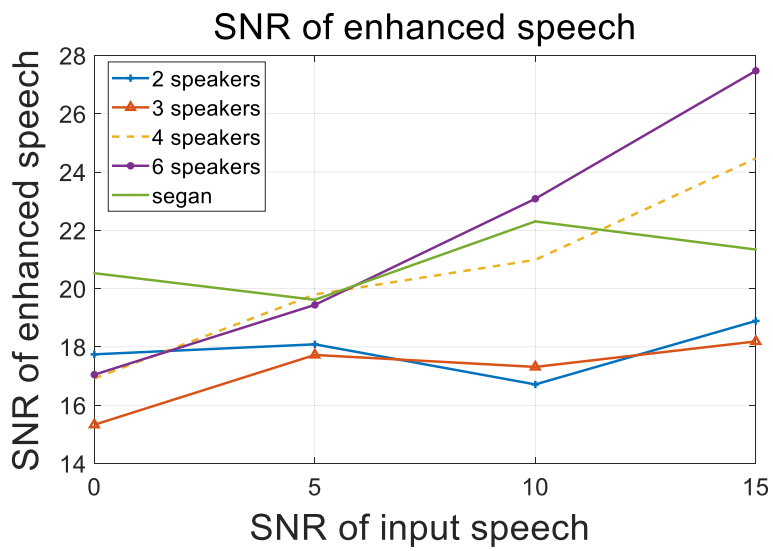


**Fig. 8** The SNR of enhancement speech. It shows the SNR of enhanced speech of test data in improved WCGAN models with different numbers of speakers and SEGAN model with the average SNR. The performance of the four-speaker model and the six-speaker model are almost the same at low SNR condition

speech but not good enough at 15 dB than the six-speaker model.

According to the two-graph information, it indicates that the four-speaker SE model has a better enhancement effect than the other models, which means that more labeled data plays a positive role to increase the enhancement and generalization performance of speech system but is not in the direct ratio when the training data is fixed.

From Table 1, we notice that improved model achieves a higher average PESQ score than that by the SEGAN model, which means added elastic work improves the intelligibility and speech quality. The PESQ improvements from 5 to 21% are observed. Although the improvement of SNR of the two-scene model is not the best, which means that it retains more voice component and obtains a highest speech quality score. The PESQ score of the four-scene model is lower than other situations but performs best in SNR on an improved model.

While the improvement in SNR of the speech is higher, relatively, the more voice component loss that causes a lower speech quality score. So, it is a significant strategy to balance the relationship between the SNR and the quality of enhanced speech. Furthermore, the

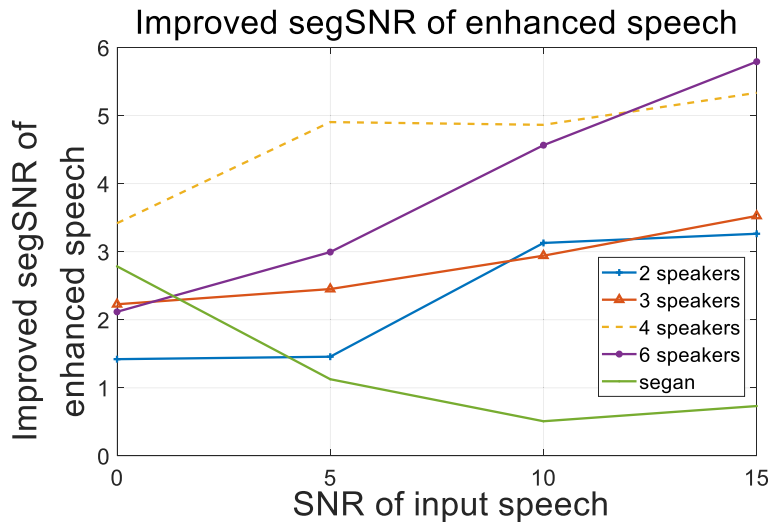


**Fig. 9** The segSNR of enhancement speech. It describes the segmented SNR of enhanced speech in different models. The four-speaker model shows a better performance than other models in low SNR of input speech but not good enough at 15 dB than the six-speaker model

**Table 1** The standard of speech quality evaluation PESQ

| Scenes PESQ | 2 | 3 | 4 | 6 | 8 | Average |
|---|---|---|---|---|---|---|
| IWCGAN | 1.915 | 1.714 | 1.497 | 1.714 | 1.634 | 1.695 |
| SEGAN | 1.908 | 1.443 | 1.423 | 1.412 | 1.855 | 1.608 |

Improved model achieves a higher average PESQ score than that by the SEGAN model, which means added elastic work improves the intelligibility and speech quality. The PESQ improvements from 5 to 21% are observed

enhancement and generalization performance of the speech system works effectively when the training data and the labeled data meet a certain proportion. We should train and design the network based on the available information of the actual data.

## 6 Conclusions

In this paper, we investigated the use of improved Wasserstein conditional GAN for speech enhancement in low-data environment. We add an elastic network to the generator loss to generalize the network performance and simplified the SE model. Experiments show that the convergence performance of the improved algorithm outperforms SEGAN in low-resource environment. The available information of input data on the system are closely related with the generalization performance of system; furthermore, we can also design the effective network structure like designing the number of layers, choosing the activity functions according to the available information of input data.

#### Abbreviations

GAN: Generative adversarial network; PESQ: Perceptual evaluation of speech quality; SE: Speech enhancement; SEGAN: Speech enhancement generative adversarial network; segSNR: Segmental signal-to-noise ratio; SNR: Signal-to-noise ratio; WCGAN: Wasserstein conditional generative adversarial network

#### Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC) (No.61671075) and Major Program of National Natural Science Foundation of China (No.61631003).

#### Funding

The funding was supported by National Natural Science Foundation of China (NSFC) (No.61671075) and Major Program of National Natural Science Foundation of China (No.61631003).

#### Authors' contributions

SQ and TJ conceptualized the idea and designed the experiments. SQ contributed in writing and draft preparation and TJ supervised the research. All authors read and approved the final manuscript.

#### Competing interests

The authors declare that they have no competing interests.

## Publisher's Note



Received: 13 April 2018 Accepted: 3 July 2018
Published online: 17 July 2018

#### References

1. Mai K V, Pastor D, Bey AAE, et al. Combined detection and estimation based on mean-square error log-spectral amplitude for speech enhancement[C]. GRETSI. Sep 2017, Juan-Les-Pins, France
2. Y Bando, K Itoyama, M Konyo, et al., Speech enhancement based on Bayesian low-rank and sparse decomposition of multichannel magnitude spectrograms[J]. IEEE/ACM Trans Audio Speech Lang Process **99**, 1 (2017)
3. Deng F, Bao C C, Jia M S, HMM-based cue parameters estimation for speech enhancement[C]. International Symposium on Chinese Spoken Language Processing. IEEE, 2017:1–4.
4. Qiao L, Zhang X, Chen X, et al, Speech enhancement using non-negative matrix factorization solved by improved alternating direction method of multipliers[C]. International Conference on Progress in Informatics and Computing IEEE, pp. 374–378 (2017)
5. Mohammed S, Tashev I. A statistical approach to semi-supervised speech enhancement with low-order non-negative matrix factorization[C]. IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE (2017).
6. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. International Conference on Neural Information Processing Systems. MIT Press. Vol.3, pp. 2672-2680 (2014)
7. CC Hsu, HT Hwang, YC Wu, et al., *Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks[J]* (2017), pp. 3364–3368
8. Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of Wasserstein GANs[J]. 2017.
9. Pascual S, Bonafonte A, Serrà J. SEGAN: speech enhancement generative adversarial network[J]. 2017.
10. Michelsanti D, Tan ZH, Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification[C]. INTERSPEECH 2017 August 20–24, Stockholm, Sweden (2017)
11. Wang Z, Li X, Wang X, et al., A DNN-HMM approach to non-negative matrix factorization based speech enhancement[C]. INTERSPEECH 2016 September 8–12, San Francisco, USA (2016)
12. Y Xu, J Du, LR Dai, et al., An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal Process Lett **21**(1), 65–68 (2014)
13. C Huemmer, A Schwarz, R Maas, et al., A new uncertainty decoding scheme for DNN-HMM hybrid systems with multichannel speech enhancement[C]. IEEE International Conference on acoustics, Speech and Signal Processing. IEEE, 5760–5764 (2016)
14. Radford A, Metz L, Chintala S, Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv: 1511.06434, (2015).
15. Kumar A, Florencio D. Speech enhancement in multiple-noise conditions using deep neural networks[J]. 2016.
16. X Lu, Y Tsao, S Matsuda, et al., Speech enhancement based on deep denoising autoencoder[C]. Interspeech., 436–440 (2013)
17. L Sun, J Du, LR Dai, et al., in *Hands-free speech communications and microphone arrays. IEEE*. Multiple-target deep learning for LSTM-RNN based speech enhancement[C] (2017)
18. W Han, X Zhang, G Min, et al., Perceptual weighting deep neural networks for single-channel speech enhancement[C]. 2016 12th World Congress on Intelligent Control and Automation, pp. 446–450 (2016)
19. C Valentini-Botinhao, X Wang, S Takaki, et al., in *ISCA Speech Synthesis Workshop*. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech[C] (2016)
20. F Weninger, H Erdogan, S Watanabe, et al., *Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR[J].*, vol 9237 (2015), pp. 91–99
21. Erdogan H, Hershey J R, Watanabe S, et al, Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks[C]. Acoustics, speech and signal processing (ICASSP), 2015 IEEE International Conference on. IEEE (2015). p. 708712.
22. KFC Yiu, KY Chan, SY Low, et al., A multi-filter system for speech enhancement under low signal-to-noise ratios[J]. J Ind Manag Optim **5**(3), 671–682 (2017)

23. Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks[J]. arXiv preprint arXiv:1701.04862 (2017).
24. Karras T, Aila T, Laine S, et al, Progressive growing of GANs for improved quality, stability, and variation[J]. arXiv preprint arXiv:1710.10196 (2017).
25. Berthelot D, Schumm T, Metz L. Began, Boundary equilibrium generative adversarial networks[J]. arXiv preprint arXiv:1703.10717 (2017).
26. Fu S W, Yu T, Lu X, et al, Raw waveform-based speech enhancement by fully convolutional networks[J] (2017).