Speech Dereverberation Based on Improved Wasserstein Generative Adversarial Networks

Lufang Rao and Junmei Yang
School of Electronic and Information, South of China Technology University, Tianhe district, Guangzhou, China
Email: 201820110598@scut.edu.cn; yjunmei@scut.edu.cn

Abstract. In reality, the sound we hear is not only disturbed by noise, but also the reverberant, whose effects are rarely taken into account. Recently, deep learning has shown great advantages in speech signal processing. But among the existing dereverberation approaches, very few methods apply deep learning at the waveform level. In addition, in the case of severe reverberation, the conventional speech dereverberation methods perform poorly, such as MCLP (multi-channel linear prediction). We proposed a new speech dereverberation method in this paper, which is based on improved WGAN (Wasserstein Generative Adversarial Networks), called WGAN-GP, whose generator uses strided-convolutional networks and the discriminator is structured on DNNs. Due to the addition of the gradient penalty item, WGAN-GP improves the stability of training and the generalization of the model. In the case of severe reverberation, according to the experimental results, the proposed system can perform better than MCLP. As the proposed method based on WAGN-GP can improve speech quality, speech signal processing systems may be able to apply it to pre-processing stage.

1. Introduction
In a room, reverberation caused by reflections as well as background noise will interfere with the source signal. The reflected signals are different delayed versions of the source speech signal. In many occasions, reverberation often brings interference, resulting in poor performance of the acoustic receiving system. For instance, in the automatic speech recognition system, strong reverberation conditions will cause a serious decrease in the accuracy [1, 2]. That is why reducing the influence of the reverberation on the sound receiving system, that is, dereverberation, is a significant topic.

In traditional speech signal processing algorithms, speech enhancement algorithm, beamforming technology algorithm and statistical model-based algorithm are classic adaptive dereverberation algorithms. Homomorphic transformation is a classic speech dereverberation method that separates clean speech from reverberation speech. An early technique is based on the finding that the impacts of reverberation exist in the linear prediction (LP) residual signal of reverberant speech. Therefore, the effect of reverberation can be eliminated by enhancing the residual signal. Griebel and Brandstein et al. [3], Gillespie et al. in [4], and Yegnanarayana et al. [5] proposed different methods to enhance the LP residual signal. And Mosayyebpour et al. [6] found that skewness is a better reverberation metric than kurtosis. Multi-Channel Linear Prediction (MCLP) [7-9], an adaptive algorithm that could achieve high quality dereverberation, also known as WPE (Weighted Prediction Error) algorithm [7, 8, 10]. Spectral subtraction is widely used in noise reduction [11], which can also be used to suppress reverberation. In recent years, deep learning has developed so rapidly and deep neural networks (DNNs) have demonstrated their powerful learning capabilities. In the field of speech signal processing, DNNs were
first used for speech denoising [12] and later in speech dereverberation [2, 13, 14]. DNNs with different networks structures [15-17] can learn a mapping function through a large amount of training data, which can be able to map a reverberant speech into a dereverberant speech. The deep learning method for dereverberation is either based on spectrum mapping [13] or masking [18].

Almost all traditional methods have high computational complexity and each step is carefully designed. In practical applications, a series of conditions must also be strictly met, so it is severely restricted. All the above DNN-based speech dereverberation methods can only work under the assumption that there is a solution of minimum mean square error (MSE) between the outputs of network and the referenced speech [19].

When the probability density cannot be calculated, some generation models that traditionally rely on the natural distribution of the data cannot be learned and applied on it. However, GANs can still be used in this case, because GANs introduces a very smart training mechanism for internal adversarial, which can approach some complex objective functions. In 2014, Goodfellow first proposed the generative adversarial networks (GANs) [20]. Arjovsky et al. [21] proposes a milestone improvement, which must be a significant breakthrough in the development of the subsequent GANs. We proposed a speech dereverberation method based on improved Wasserstein generative adversarial networks (WGAN-GP) in our study. It works at waveform level, there is no need to extract time-frequency features like DNN, and so the dereverberation process is faster. In addition, WGAN-GP has been proved that the training process is more stable, so it has stronger generalization ability. The experimental results also prove that in the case of serious reverberation, it can also perform better than MCLP.

2. Improved Wasserstein Adversarial Generative Networks

Generator and discriminator are two models in GANs. In the training process, the purpose of discriminator D is to try to distinguish the signal output by G from the real sample, and the purpose of generator G is to output a signal close to the real sample to deceive discriminator D. The two models compete to improve their abilities during the training process and the final generator can output the samples that D cannot distinguish. $G(z)$ denotes the data that G maps the input $z$ from noise distribution $p_z(z)$ into (i.e., generating a picture). $D(x)$ denotes the probability that $x$ is obtained from the real training data distribution $P_{data}(x)$ instead of $P_{G(z)}$. Accordingly, researchers often defined the following formula as optimized objective function of GANs:

$$
\min_G \max_D V(D,G) = E_{x \sim P_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))]
$$

(1)

In view of the uncontrollable shortcomings of GANs itself, the authors proposed to add supervision information $x_s$ to guide the GANs network to generate [22].

$$
\min_G \max_D V(D,G) = E_{x \sim P_{data}(x_s)}[\log D(x, x_s)] + E_{z \sim p_z(z), x \sim P_{data}(x)}[\log(1 - D(G(z), x_s))]
$$

(2)

The gradient vanishing is always a troublesome problem in the training of deep neural network models. In addition, mode collapse and unstable training are also weakness of GANs. The authors proposed WGANs in [21], which worked out solutions to the above shortcomings of GANs. In addition, there is no need to pay special attention to the training balance of generator and discriminator. Finally, they proposed a specific loss function to guide the training progress of WGANs. The smaller this value is, the better the GANs training.

The difference is that discriminator model in WGANs is no longer to calculate the probability for binary classification, but to compute the Wasserstein Distance (WD) of the real training data distribution $P_{data}(x)$ and generator’s distribution $P_{G(z)}$. The smaller the value is, the better. WD is defined as follows:
In the above formula, $K$ is a constant, $\|f_w(x)\|_{\text{Lipschitz}} \leq K$ represents that Lipschitz constant of the function $f_w(x)$ is less than $K$, $w$ is training parameter, and $x$ is from real data distribution. The discriminator’s and generator’s loss function of WGANs are as follows:

$$\min L(D) = E_{z \sim \mathcal{P}_{\text{data}}} [f_w(G(z))] - E_{x \sim \mathcal{P}_{\text{data}}} [f_w(x)]$$

$$\min L(G) = E_{z \sim \mathcal{P}_{\text{data}}} [f_w(G(z))]$$

Although the WGANs described above can solve the instability problem of the GANs model training, the weight clipping of the parameters will lead to optimization difficulties. Because of the limited range of weights, neural network architectures that attempt to obtain the maximum gradient norm often end up with a simple learning function. Fortunately, the authors immediately proposed gradient penalty (GP) to solve the above problem of WGANs in [23]. With a gradient penalty term, the equation 5 can be redefined as:

$$\min L(D) = E_{z \sim \mathcal{P}_{\text{data}}} [f_w(x)] - E_{x \sim \mathcal{P}_{\text{data}}} [f_w(G(z))]$$

$$+ \lambda (\|\nabla_{x \sim \mathcal{P}_{\text{data}}} [f_w(x)]\| - 1)^2$$

where $\lambda$ is a penalty coefficient, and the $x_t = \varepsilon x_t + (1-\varepsilon)x_g$, where $x_t \sim \mathcal{P}_{\text{data}}(x)$, $x_g \sim \mathcal{P}_{G(z)}$, $\varepsilon \sim \text{Uniform}[0,1]$.

3. Speech Dereverberation Based on WGAN-GP

In the proposed method, we construct generator with fully convolutional neural networks (FCNNs), which is structured similarly to an encoder-decoder. The encoder is made up by some strided convolutional layers, and the activation layer of each strided convolution layer is parametric rectified linear units (PReLUs) [24]. Reversely, decoder decompresses the signal that concatenated the vector h compressed by encoder with the noise vector z which improves the robustness of the network. In this paper, we obtain noise vector z from Gaussian distribution with a spherical structure. After the encoder and decoder, the output of generator is the dereverberant speech $\hat{x} = G(\tilde{x})$. Because we use fewer pooling layers than [24], more of the speech signal characteristics can be reserved, theoretically we can get higher speech quality [25].

The discriminator of WGAN-GP is no longer to solve a binary classification problem, but to fit the Wasserstein distance approximately, so the binary classification problem becomes a regression task. Although there are more training parameters in DNNs than CNNs, the computational complexity is smaller [25]. The generator’s and discriminator’s architecture are respectively shown in figures 1 and 2.

As suggested in Ref. [24], the $L_1$ norm term, which can avoid overfitting, is appended to the generator’s loss function of WGAN-GP. In additional, the reverberant speech is added as the extra information, so in the end, the following formulas are generator’s and discriminator’s loss functions of WGAN-GP we used respectively:

$$\min L(G) = -E_{z \sim \mathcal{P}_{\text{data}}(z)\cdot \mathcal{P}_{\text{Gauss}}(\tilde{x})} [f_w(G(z, \tilde{x}), \tilde{x})]$$

$$+ \lambda (\|G(z, \tilde{x}) - x\|)$$

$$\min L(D) = E_{z \sim \mathcal{P}_{\text{data}}(z)\cdot \mathcal{P}_{\text{Gauss}}(\tilde{x})} [f_w(G(z, \tilde{x}), \tilde{x})] -$$

$$E_{x \sim \mathcal{P}_{\text{data}}(x)\cdot \mathcal{P}_{\text{Gauss}}(\tilde{x})} [f_w(x, \tilde{x}) + \lambda (\|\nabla_{x \sim \mathcal{P}_{\text{data}}(x)\cdot \mathcal{P}_{\text{Gauss}}(\tilde{x})} [f_w(x, \tilde{x})]\| - 1)^2$$
4. Experiments

4.1. Experimental Setup

The data set utterances of experiment consisted of 30 different speakers, which are all from the Voice Bank corpus [26]. The reverberation speech is a linear superposition of the source speech and its different delayed reflection signals, so we can make a convolution operation of the original clean speech $x(t)$ and the room impulse response (RIR) $h(t)$ to generate reverberant speech $\tilde{x}(t)$. We utilize a generator of room impulse response [27, 28] to obtain various RIRs. Different RIRs can be obtained by changing the relative location of the receiver and the source speech signal. In the experiment, we keep the distance from the source speech signal to the receiver constant, which ensures that the direct speech signal ratio is basically unchanged [29]. For each of 10 different relative positions, we have generated 3 different RIRs ($T_{60}$ is 0.3 s, 0.6 s and 0.9 s). Five values of $T_{60}$ at 0.5 s, 0.6 s, 0.7 s, 0.8 s and 0.9 s for testing. In summary, we have $10 \times 3 \times (30)$ RIRs for the 11572 train speeches, 5 RIRs for the 824 test speeches.

As shown in figure 3, the training sample of the discriminator consists of two parts: we add the reverberant speech as extra information to the clean speech and the dereverberant speech, as real samples and fake samples, respectively. We trained the networks model with SGD for discriminator and Adam for generator. Follow the suggestions of the training GAN that the researchers have summarized, we utilize Two Time-Scale Update Rule (TTUR) [30], so we set generator’s learning rate to 0.0002 and discriminator’s learning rate to 0.0004. The above two tricks make the discriminator is trained more, which can deliver a correct training direction to the generator. As suggested in the [23], the number of updates to be applied to D before G is set to 2, and we set the coefficient of the gradient penalty term $\lambda$ to 10. Before training, we transferred the sampling frequency of original utterances from 48 kHz to 16 kHz.
In both train and test phase, we use Dropouts in G and set the probability to 0.5, providing the noise by Dropouts to make the generator more generalizable. The discriminator uses Layer Normalization as the authors in [23] recommended.

4.2. Evaluation Results

We compare our proposed method with MCLP from following subjective and objective metrics, we compute these measures (the higher the better) at different $T_{60}$ (The larger the value, the more serious reverberation). Short-time-objective-intelligibility-measure (STOI) is a function that outputs a scalar value ranging from 0 to 1. The bigger this value, the better the speech intelligibility [31]. Mean opinion score (MOS) is a subjective metric, which evaluates the speech signal quality as a whole (from 1 to 5) [32]. We predicted the subjective measurement from two aspects: signal distortion (CSIG) and overall effect (COVL).

The results of our experiment are displayed in tables 1-3. According to the CSIG scores, the MCLP approach is effective only at $T_{60} = 0.5T_s$, and in the case of longer reverberation times, it even shows a drop in speech quality, the WGAN-GP models can still maintain a good dereverberation effect. The scores of the STOI and COVL shows that MCLP approach can provide limited improvement with the increasement of reverberation time. Significant improvements over the MCLP approach are observed, the WGAN-GP model’s average STOI, CSIG, and COVL score also respectively increase by 2.1%, 16%, and 15.1%, which illustrates that our proposed method can achieve higher speech quality without reducing short-time-objective-intelligibility.

**Table 1. The scores of STOI (in %).**

| $T_{60}$ (s) | 0.5  | 0.6  | 0.7  | 0.8  | 0.9  | Avg.  |
|--------------|------|------|------|------|------|-------|
| Models       |      |      |      |      |      |       |
| Reverberant  | 76.82| 72.32| 69.64| 68.78| 65.85| 70.69 |
| MCLP         | 79.44| 75.06| 71.85| 71.16| 68.27| 73.16 |
| WGAN-GP      | 79.29| 76.42| 73.69| 73.14| 70.67| 74.65 |

**Table 2. The scores of CSIG.**

| $T_{60}$ (s) | 0.5  | 0.6  | 0.7  | 0.8  | 0.9  | Avg.  |
|--------------|------|------|------|------|------|-------|
| Models       |      |      |      |      |      |       |
| Reverberant  | 3.2214| 2.8106| 2.5779| 2.4145| 2.3946| 2.6840|
| MCLP         | 3.3306| 2.7739| 2.2415| 2.0482| 2.1119| 2.5017|
| WGAN-GP      | 3.1596| 2.9871| 2.9006| 2.8202| 2.7872| 2.9311|
Table 3. The scores of COVL.

| Models      | 0.5   | 0.6   | 0.7   | 0.8   | 0.9   | Avg.  |
|-------------|-------|-------|-------|-------|-------|-------|
| Reverberant | 2.4037| 2.1541| 2.0037| 1.7419| 1.7712| 1.7538|
| MCLP        | 2.4536| 2.2329| 2.0021| 1.7668| 1.7858| 2.0486|
| WGAN-GP     | 2.5615| 2.4237| 2.3171| 2.1366| 2.1297| 2.3139|

The above metrics measure the method from the perspective of the time-domain. We also proved the effectiveness from the frequency-domain perspective, we randomly selected a most severe reverberant speech from the test set. The spectrogram of the reverberant speech is shown in figure 4a. After processed separately by MCLP approach and our approach, we obtained spectrogram of the dereverberated speech, they are shown in figures 4b and 4c. Obviously, the proposed method can remove more smearing effects caused by reverberation than MCLP.

![Figure 4](image)

(a) Reverberant speech   (b) MCLP dereverberant speech   (c) WGAN-GP dereverberant speech

**Figure 4.** The reverberant speech spectrogram and corresponding spectrogram of the dereverberant speech processed by the MCLP approach and our approach.

5. Conclusion

A speech dereverberation method is proposed in this paper, which is based on WGAN-GP to perform reverberation suppression. Our model’s generator with FCNN is used for dereverberation and discriminator with DNN computes WD, which can be fed back to the generator to indicate the training direction of the generator. The main research contents of this paper are followings: (1) A new speech dereverberation method based on WGAN-GP is introduced, and the effectiveness of speech dereverberation is proved by the experimental results; (2) The problem of poor performance at long reverberation times of traditional dereverberation methods is solved. Proved by experimental results, the proposed method can obtain better performance of dereverberation than MCLP at T-F domain, and it has strong generalization in practical applications. According to the experimental results, in future research, we may be able to combine traditional methods. For reverberant speech, first use WGAN-GP for pre-processing, then use MCLP method, so that it may get better dereverberation effect.

References

[1] Kinoshita K, et al. 2013 The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.* pp 1-4.

[2] Kinoshita K, et al. 2016 A summary of the reverb challenge: State-of-the-art and remaining challenges in reverberant speech processing research *EURASIP J. Adv. Signal Process.* 7 (1) 1-19.

[3] Griebel S and Brandstein M 1999 Wavelet Transform Extrema Clustering For Multi-Channel Speech Dereverberation *IEEE Workshop on Acoustic Echo and Noise Control.*
[4] Gillespie B W, Malvar H S and Florencio D A 2001 Speech dereverberation via maximum-kurtosis subband adaptive filtering Proc. ICASSP. IEEE Int. Conf. Acoust., Speech, Signal Process (Salt Lake City, UT, USA) pp 3701-3704.

[5] Yegnanarayana B, Prasanna S R M and Rao K S 2002 Speech enhancement using excitation source information IEEE Int. Conf. Acoust., Speech, Signal Process (Orlando, FL) pp I-541-I-544.

[6] Mosayyebpour S, Esmaeili M and Gulliver T A 2013 Single-microphone early and late reverberation suppression in noisy speech IEEE Trans. Audio Speech Lang. Process 21 (2) 322-335.

[7] Nakatani T, Yoshioka T, Kinoshita K, Miyoshi M and Juang B 2008 Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation IEEE Int. Conf. Acoust., Speech, Signal Process (Las Vegas, NV) pp 85-88.

[8] Nakatani T, Yoshioka T, Kinoshita K, Miyoshi M and Juang B 2010 Speech dereverberation based on variance-normalized delayed linear prediction IEEE Trans. Audio, Speech, Lang. Process 18 (7) 1717-1731.

[9] Jukić A, van Waterschoot T, Gerkmann T and Doclo S 2015 Multi-channel linear prediction-based speech dereverberation with sparse priors IEEE/ACM Trans. Audio, Speech, Lang. Process 23 (9) 1509-1520.

[10] Yoshioka T, Nakatani T, Kinoshita K, Miyoshi M, et al. 2011 Blind separation and dereverberation of speech mixtures by joint optimization IEEE Trans. Audio, Speech, Lang. Process 19 (1) 69-84.

[11] Naylor P A and Gaubitch N D 2010 Speech Dereverberation (London, U.K.: Springer).

[12] Xu Y, Du J, Dai L-R and Lee C-H 2014 An experimental study on speech enhancement based on deep neural networks IEEE Signal Processing Letters 21 (1) 65-68.

[13] Wu B, Li K, Yang M, Lee C-H, et al. 2017 A reverberation-time-aware approach to speech dereverberation based on deep neural networks IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 25 (1) 102-111.

[14] Kinoshita K, Delcroix M, Kwon H, Mori T and Nakatani T 2017 Neural network-based spectrum estimation for online WPE dereverberation Interspeech 384-388.

[15] Han K, et al. 2015 Learning spectral mapping for speech dereverberation and denoising IEEE/ACM Trans. Audio Speech Lang. Process. 23 (6) 982-992.

[16] Weninger F, Watanabe S, Tachioka Y and Schuller B 2014 Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp 4623-4627.

[17] Park S R and Lee J 2016 A fully convolutional neural network for speech enhancement arXiv preprint arXiv:1609.07132.

[18] Williamson D and Wang D 2017 Time-frequency masking in the complex domain for speech dereverberation and denoising IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[19] Wang K, Zhang J, Sun S, et al. 2018 Investigating generative adversarial networks based speech dereverberation for robust speech recognition arXiv preprint arXiv:1803.10132v3.

[20] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, WardeFarley D, Ozair S, Courville A and Bengio Y 2014 Generative adversarial nets Advances in Neural Information Processing Systems (NIPS) 2672-2680.

[21] Arjovsky M, Chintala S and Bottou L 2017 Wasserstein GAN Proceedings of the 34th International Conference on Machine Learning pp 214-223.

[22] Isola P, Zhu J-Y, Zhou T and Efros A A 2016 Image-to-image translation with conditional adversarial networks arXiv preprint arXiv: 1611.07004.

[23] Gulrajani I, Ahmed F, Arjovsky M, et al. 2017 Improved training of wasserstein GANs Advances in Neural Information Processing Systems 30 5767-5777.
[24] Pascual S, Bonafonte A and Serr’a J, 2017 Segan: Speech enhancement generative adversarial network arXiv preprint arXiv:1703.09452.

[25] Ye S, Jiang T, Qin S, Zou W and Deng C 2018 Speech enhancement based on a new architecture of wasserstein generative adversarial networks 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP) (Taipei City) pp 399-403.

[26] Veaux C, Yamagishi J and King S 2013 The voice bank corpus: Design, collection and data analysis of a large regional accent speech database Int. Conf. Oriental COCOSDA, held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE) pp 1-4.

[27] Habets E 2014 Room impulse response generator https://www.audiolabserlangen.de/fau/professor/habets/software/rir-generator.

[28] Allen J B and Berkley D A 1979 Image method for efficiently simulating small-room acoustics J. Acoustical Soc. Amer. 65 943-950.

[29] Zhao Y, Wang Z and Wang D 2019 Two-stage deep learning for noisy-reverberant speech enhancement IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (1) 53-62.

[30] Martin H, Hubert R, Thomas U and Bernhard N 2017 GANs trained by a two time-scale update rule converge to a local Nash equilibrium arXiv preprint arXiv:1706.08500.

[31] Taal C H, Hendriks R C, Heusdens R and Jensen J 2011 An algorithm for intelligibility prediction of time-frequency weighted noisy speech IEEE Trans. Audio, Speech, Lang. Process. 19 (7) 2125-2136.

[32] Hu Y and Loizou P C 2008 Evaluation of objective quality measures for speech enhancement IEEE Trans. on Audio, Speech, and Language Processing 16 (1) 229-238.