Double Adversarial Network based Monaural Speech Enhancement for Robust Speech Recognition

Zhihao Du[1], *Jiqing Han*[1], *Xueliang Zhang*[2]

[1]School of Computer Science and Technology, Harbin Institute of Technology, China
[2]Department of Computer Science, Inner Mongolia University, China

{duzhihao, jqhan}@hit.edu.cn, cszxl@imu.edu.cn

Abstract

To improve the noise robustness of automatic speech recognition (ASR), the generative adversarial network (GAN) based enhancement methods are employed as the front-end processing, which comprise a single adversarial process of an enhancement model and a discriminator. In this single adversarial process, the discriminator is encouraged to find differences between the enhanced and clean speeches, but the distribution of clean speeches is ignored. In this paper, we propose a double adversarial network (DAN) by adding another adversarial generation process (AGP), which forces the discriminator not only to find the differences but also to model the distribution. Furthermore, a functional mean square error (f-MSE) is proposed to utilize the representations learned by the discriminator. Experimental results reveal that AGP and f-MSE are crucial for the enhancement performance on ASR task, which are missed in previous GAN-based methods. Specifically, our DAN achieves 13.00% relative word error rate improvements over the noisy speeches on the test set of CHiME-2, which outperforms several recent GAN-based enhancement methods significantly.

Index Terms: speech enhancement, adversarial training, speech recognition, CHiME-2

1. Introduction

Monaural speech enhancement aims at separating clean speeches from the noisy backgrounds by using a single microphone. Since monaural speech enhancement is formulated as a supervised learning problem [1], many deep learning techniques are introduced into this community, such as the convolutional and recurrent networks [2, 3, 4].

In recent years, the generative adversarial network (GAN) has been widely researched in the machine learning community, which is optimized through an adversarial training process [5]. In the GAN-based speech enhancement methods, a discriminator is added to perform the adversarial training with the enhancement model. Early GAN-based models, such as the SEGAN [6], perform better than the conventional methods but worse than their supervised counterparts [7] in terms of short term objective intelligibility (STOI) [8] and perceptual evaluation of speech quality (PESQ) [9]. Therefore, different loss functions are explored for GAN-based enhancement methods in [7]. By employing the relativistic adversarial loss [10], GAN-based methods eventually achieve higher speech intelligibility and perceptual quality than the supervised methods [11].

Inspired by the progress in improving the speech intelligibility and perceptual quality, recent studies try to improve the noise (including reverberation) robustness of automatic speech recognition (ASR) by adopting a GAN-based enhancement method as the front-end processing. For the reverberation noises, GAN-based methods have improved the recognition performance of a multi-conditional trained (MCT) ASR system [12]. For the additive noises, the GAN-based enhancement method is explored on the logarithm compressed fbank (log-fbank) domain, which reduces the word error rate (WER) of a clean-trained ASR system [13]. However, this method fails to improve the recognition performance of a MCT ASR system without retraining the acoustic model [13]. To improve the noise robustness of the MCT ASR system, an additional phoneme classifier is involved into the adversarial enhancement process, resulting in the multi-target learning scheme [14, 15].

Current GAN-based enhancement methods for robust ASR comprise a single adversarial process, which encourages the discriminator to learn the representations maximizing the distance between enhanced and clean speeches rather than modeling the distribution of clean speeches. However, without learning the distribution, the representations learned by the discriminator are less meaningful, which may mislead the enhancement model, leading to the recognition performance degradation. To overcome this problem, we propose the double adversarial networks (DANs), which consist of the adversarial generation process (AGP) and the adversarial enhancement process (AEP). In AGP, an additional generator is involved, and the discriminator is trained to model the speech distribution by playing the min-max game with the generator. In AEP, the learned information of speech distribution is propagated from the discriminator to the enhancement model through an adversarial training process. To utilize the learned representations, we further propose a functional mean square error (f-MSE), in which the MSE is calculated on a meaningful and distinguishable feature domain defined by the discriminator.

2. Double adversarial networks

Double adversarial networks (DANs) consist of three components, i.e. the discriminator $\mathbf{D}$, the generator $\mathbf{G}$ and the enhancement model $\mathbf{E}$. These three components play the min-max games in two adversarial processes, i.e. AEP and AGP. An overview of the proposed DAN is shown in Figure 1.

2.1. Adversarial enhancement process

In AEP, the primary task of enhancement model $\mathbf{E}$ is to reconstruct the clean speech $\boldsymbol{s}$ from the noisy one $\boldsymbol{x}$, and the discriminator $\mathbf{D}$ is trained to distinguish the real clean speech $\boldsymbol{s}$ from the enhanced speech $\hat{\boldsymbol{s}}$. By taking an adversarial training between $\mathbf{E}$ and $\mathbf{D}$, the enhancement model tries to fool the discriminator by producing speeches that are similar to their real clean

This research was supported by National Key Research and Development Program of China under Grant 2017YFB1002102, National Natural Science Foundation of China under Grant U1736210.

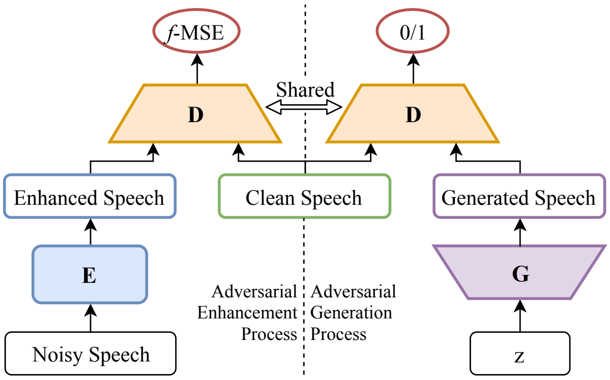

Figure 1: *The framework of double adversarial networks.*

counterparts, and the discriminator is encouraged to learn the representations that maximize the distance between enhanced and clean speeches. The original adversarial loss function [5] is prone to cause the problem of vanishing gradient in the early stages of training, which leads to the model collapse and unstable training process [16]. In this paper, we employ the least squares loss functions [17], which are similar to [6, 7]. The adversarial loss functions of $\mathbf{D}$ and $\mathbf{E}$ are given as follows:

$$\begin{aligned}\mathcal{L}_{D(E)} &= \frac{1}{2}\mathbb{E}_{\boldsymbol{s}\sim p_{clean}}[\mathbf{D}(\boldsymbol{s})-1]^2 \\ &+ \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{noisy}}[\mathbf{D}(\mathbf{E}(\boldsymbol{x}))-0]^2 \qquad (1)\\ \mathcal{L}_E &= \mathbb{E}_{(\boldsymbol{s},\boldsymbol{x})\sim(p_{clean},p_{noisy})}\|\boldsymbol{s}-\mathbf{E}(\boldsymbol{x})]\|_2^2 \\ &+ \lambda\mathbb{E}_{\boldsymbol{x}\sim p_{noisy}}[\mathbf{D}(\mathbf{E}(\boldsymbol{x}))-1]^2 \qquad (2)\end{aligned}$$

where λ is a hyper-parameter to balance the reconstruction and adversarial losses.

The second term in (2) aims at fooling the discriminator. However, the task of enhancement model is to reconstruct the clean speech rather than simply fool the discriminator. Therefore, a functional mean square error, f-MSE, is proposed to replace the adversarial loss term. f-MSE evaluates the similarity of enhanced and clean speeches in a more distinguishable feature domain defined by the discriminator $\mathbf{D}$. The modified loss function for $\mathbf{E}$ is defined as follows:

$$\begin{aligned}\mathcal{L}_E &= \mathbb{E}_{(\boldsymbol{s},\boldsymbol{x})\sim(p_{clean},p_{noisy})}\|\boldsymbol{s}-\mathbf{E}(\boldsymbol{x})]\|_2^2 \\ &+ \lambda\mathbb{E}_{(\boldsymbol{s},\boldsymbol{x})\sim(p_{clean},p_{noisy})}[\mathbf{D}(\boldsymbol{s})-\mathbf{D}(\mathbf{E}(\boldsymbol{x}))]^2 \qquad (3)\end{aligned}$$

There are three advantages of using f-MSE. First, it explicitly involves the clean speech information into the loss term. Second, this loss term still maintains the adversarial relationship between $\mathbf{E}$ and $\mathbf{D}$, therefore, it can also benefit from the adversarial training. Third, f-MSE is calculated on a more distinguishable feature domain, which is learned and changed during the adversarial training process.

2.2. Adversarial generation process

In AGP, an additional generator $\mathbf{G}$ is added and trained against the discriminator. By taking the adversarial training between $\mathbf{G}$ and $\mathbf{D}$, the discriminator is forced to learn not only the differences between clean and enhanced speeches but also the distribution of clean speeches. The loss functions for $\mathbf{D}$ and $\mathbf{G}$ are given as follows:

$$\begin{aligned}\mathcal{L}_{D(G)} &= \frac{1}{2}\mathbb{E}_{\boldsymbol{s}\sim p_{clean}}[\mathbf{D}(\boldsymbol{s})-1]^2 \\ &+ \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_z}[\mathbf{D}(\mathbf{G}(\boldsymbol{z}))-0]^2 \qquad (4)\\ \mathcal{L}_G &= \mathbb{E}_{\boldsymbol{z}\sim p_z}[\mathbf{D}(\mathbf{G}(\boldsymbol{z}))-1]^2 \qquad (5)\end{aligned}$$

where $\boldsymbol{z}$ is a d-dimensional random vector following the standard Gaussian distribution $\mathcal{N}(0,1)$.

2.3. Stabilizing the training process

GANs always suffer the instability during their training process [18, 19]. In this paper, we employ the gradient-penalty (GP) regularization [18] and a tuned update strategy to stabilize the training process.

The GP regularization penalizes the discriminator if the outputs have too large gradients with respect to the inputs:

$$\mathcal{L}_{D(GP)} = \mathbb{E}_{\boldsymbol{y}\sim p_y}\left[\|\nabla_{\boldsymbol{y}}\mathbf{D}(\boldsymbol{y})\|_2 - 1\right]^2 \qquad (6)$$

where $\boldsymbol{y}$ is sampled from the convex combinations of the real clean and fake speeches: $\boldsymbol{y} = \epsilon\boldsymbol{s} + (1-\epsilon)\tilde{\boldsymbol{s}}$. ϵ is a scalar sampled from the uniform distribution $[0,1]$. In DAN, both the enhanced speeches $\hat{\boldsymbol{s}}$ and the generated speeches $\breve{\boldsymbol{s}}$ are treated as the fake samples: $\tilde{\boldsymbol{s}} = \hat{\boldsymbol{s}} \cup \breve{\boldsymbol{s}}$.

To further stabilize the training process, we update the models with a tuned strategy. Specifically, in a mini-batch, we first update the discriminator five times, then the enhancement model and the generator are updated one time. A similar unbalanced update strategy is also employed in [18].

By integrating the two adversarial processes and the gradient-penalty regularization, the loss function for the discriminator in DANs is obtained:

$$\mathcal{L}_D = \mathcal{L}_{D(E)} + \mathcal{L}_{D(G)} + \gamma\mathcal{L}_{D(GP)} \qquad (7)$$

where γ is a hyper-parameter to control the GP regularization. The loss functions for the enhancement model and the generator are given in (3) and (5), respectively.

3. Experiments

3.1. Dataset

We evaluate the proposed methods on the CHiME-2 dataset (track 2) [20]. In the training set, there are 7138 clean utterances from the WSJ0 SI-84 training set, which are recorded by 83 speakers. To simulate a noisy reverberant environment, each clean utterance is first filtered with a fixed Room Impulse Response (RIR) corresponding to a frontal position at a distance of 2 m. Then, each reverberated utterance is mixed with a random slice from the living room noise recording. The SNR level of each mixture is randomly selected from -6, -3, 0, 3, 6, 9 dB. To enrich the training data, an additive noisy environment is simulated without filtering the utterances with RIRs. In this data augmentation (DA), a seven hours noise recording provided by CHiME-2 is used to obtain the noisy mixtures, and the clean utterances and SNR levels are the same as the training set.

The development and test sets are obtained in the same manner as the training set, but the RIRs and noise slices are different from the training set. The SNR levels for development and test sets are the same as the training set. As a result, there are 2,460 noisy utterances from 10 other speakers in the development set, and the test set comprises 1,980 noisy utterances from 12 other speakers.

3.2. ASR system

The ASR system is implemented by using the Kaldi toolkit [21]. Following the Kaldi's recipe for CHiME-2, we extract the mel-frequency cepstral coefficients and built up a Gaussian mixture model-hidden Markov model system to perform the force-alignment. The aligned phoneme labels and log-fbank features are used to train a DNN-based acoustic model (AM), which is initialized with the restricted Boltzmann machine based pre-training. This AM comprises 7 hidden layers with the sigmoid activation function, and there are 2,048 units in each layer. The input of AM is a stacked 11-frame log-fbank feature centered at the current frame, and the desired output is the posterior probability of 1,960 state-clustered triphone classes. After four iterations of alignment and retraining, the AM is further optimized by the state-level minimizing Bayes risk criterion [22]. The WSJ 5k trigram language model is employed, and the weighted finite-state transducers is used for decoding.

In the official guideline provided by CHiME-2, the ASR system is trained with the noisy utterances only, but we find that such noisy-speech-only training strategy leads to an unacceptable WER for clean utterances (about 90%). Therefore, we train another ASR system with the clean and noisy utterances, resulting in a MCT system, which has much lower WERs for both clean and noisy utterances than the official one. In the following experiments, we employ this MCT ASR system as the baseline to evaluate the proposed DANs.

3.3. Model settings

A convolutional recurrent network (CRN) [4] is employed as the enhancement model $\mathbf{E}$, which is fed with the 40-dimensional log-fbank features of noisy speeches. The architecture details are given in Table 1. Batch normalization [23] and exponential linear units (ELUs) [24] are added after each convolution and deconvolution layers except the last layer, which is followed by the sigmoid function only. In addition, there is a skip connection between each convolution layer and its corresponding deconvolutional counterpart. A widely used training target, the ideal ratio mask (IRM) [25], is employed as the expected output of the enhancement model, which is calculated on the fbank domain. By multiplying the predicted IRM and noisy speech, the enhanced speech is obtained: $\hat{\boldsymbol{s}} = \boldsymbol{x} \odot \mathbf{E}(\boldsymbol{x})$, where $\odot$ represents the element-wise multiplication.

For the discriminator and generator, the architecture of DCGAN is adopted, which is stable during the training process [16]. The log-fbank slices of 40 frames are randomly clipped from the enhanced and clean features, which are treated as the fake and real samples, respectively. To match the architecture of discriminator, the log-fbank slices are upsampled to 64×64 by repeating the nearest neighbors. For the generator, a 128-dimensional Gaussian random vector is employed as the input, and the desired output is the upsampled log-fbank slices with the size of 64×64, which are also treated as the fake samples. The features of each utterance are limited to $[-1, 1]$ by the min-max normalization. To match the value limitation, the tanh nonlinearity is adopted in the output layer of the generator. The Adam optimizer [26] is used to train the models with the learning rate of 2×10^{-4} and the betas of $(0.5, 0.999)$. The utterance-level batch size for $\mathbf{E}$ is 16, and the slice-level batch size for $\mathbf{G}$ and $\mathbf{D}$ is 64. λ and γ are set to 1 and 10, respectively. More details can be found in our open source code[1]. The best model is selected by cross validation on the development set.

Table 1: *The architecture of the enhancement model. Here T denotes the number of time frames in the log-fbank features.*

layer name	input size	kernel, stride	output size
reshape_1	$T \times 40$	-	$1 \times T \times 40$
conv2d_1	$1 \times T \times 40$	$3 \times 4, (1, 2)$	$16 \times T \times 20$
conv2d_2	$16 \times T \times 20$	$3 \times 4, (1, 2)$	$32 \times T \times 10$
conv2d_3	$32 \times T \times 10$	$3 \times 4, (1, 2)$	$64 \times T \times 5$
conv2d_4	$64 \times T \times 5$	$3 \times 4, (1, 2)$	$128 \times T \times 2$
conv2d_5	$128 \times T \times 2$	$1 \times 2, (1, 1)$	$256 \times T \times 1$
reshape_2	$256 \times T \times 1$	-	$T \times 256$
lstm_1	$T \times 256$	256×1024	$T \times 1024$
lstm_2	$T \times 1024$	1024×1024	$T \times 1024$
fc	$T \times 1024$	1024×256	$T \times 256$
reshape_3	$T \times 256$	-	$256 \times T \times 1$
deconv2d_5	$512 \times T \times 1$	$1 \times 2, (1, 1)$	$128 \times T \times 2$
deconv2d_4	$256 \times T \times 2$	$3 \times 4, (1, 2)$	$64 \times T \times 5$
deconv2d_3	$128 \times T \times 5$	$3 \times 4, (1, 2)$	$32 \times T \times 10$
deconv2d_2	$64 \times T \times 10$	$3 \times 4, (1, 2)$	$16 \times T \times 20$
deconv2d_1	$32 \times T \times 20$	$3 \times 4, (1, 2)$	$1 \times T \times 40$
reshape_4	$1 \times T \times 40$	-	$T \times 40$

3.4. Compared methods

We compare the proposed DAN with four recent GAN-based speech enhancement methods. The first method is SEGAN [6], which enhances the noisy speech in the waveform domain directly. In this paper, the same model architectures and training strategies as [6] are implemented. The second one is SERGAN [11], which is based on the relativistic average loss functions [10]. In the original SERGAN, a fully convolutional U-Net architecture is employed to enhance the waveform directly. Preliminary experiments show that our used CRN in the log-fbank domain outperforms the U-Net model in the waveform domain in terms of WER. Therefore, we also implement another SERGAN, in which our used models ($\mathbf{E}$ and $\mathbf{D}$) are trained by minimizing the relativistic average loss functions [10]. This model is denoted as SERGAN-fbank. The third method is based on a conditional GAN, where the pairs of desired clean masks and noisy features are treated as real samples, and the pairs of predicted masks and noisy features are fake samples [7]. This method is denoted as MaskCoGAN. The fourth one is a GAN-based feature mapping model, which improves the recognition performance in the reverberation environment [12]. It is denoted as DereverbGAN.

For fair comparison, we replace the enhancement models in MaskCoGAN and DereverbGAN with our used CRN, since it has been verified that CRNs achieve better performance [4, 27]. All methods are trained with the same training set (including data augmentation), and all the models with the similar architecture are initialized with the same weights.

4. Results and Discussion

We first compare the proposed DAN with other methods, then the effect of each module is evaluated. Finally, we explore the outputs of $\mathbf{D}$, $\mathbf{G}$ and $\mathbf{E}$ through the training process.

4.1. Model comparison

Table 2 shows the WER results of DAN and other GAN-based methods under different SNR levels on the test set. From the table, we can see that the recognition performance cannot be improved by employing the SEGAN or SERGAN as the front-end processing, which enhance the noisy speeches on the waveform domain directly. By changing the enhancement domain from

[1] Available at https://github.com/ZhihaoDU/du2020dan.

Table 2: *The WER results of DAN and other GAN-based methods under different SNR levels on the test set.*

Models	**-6dB**	**-3dB**	**0dB**	**3dB**	**6dB**	**9dB**	**Avg.**
Baseline (MCT)	43.94	33.15	24.80	18.16	14.42	**11.38**	24.31
SEGAN [6]	74.88	68.01	60.13	48.17	40.18	33.60	54.16
SERGAN [11]	73.61	66.48	57.96	47.66	40.30	32.23	53.04
SERGAN-fbank	37.90	28.42	22.28	18.32	14.10	12.85	22.31
MaskCoGAN [7]	37.34	28.56	22.16	18.85	14.98	13.05	22.49
DereverbGAN [12]	41.21	31.30	24.75	20.73	16.39	14.53	24.82
DAN (ours)	**36.35**	**27.12**	**20.74**	**16.84**	**13.69**	12.16	**21.15**

Table 3: *The impact of each module in terms of WER on the development (Dev.) and test sets, where × and √ indicate excluding or including the sub-module.*

Models	**AGP**	**AEP**	***f*-MSE**	**DA**	**Dev. (%)**	**Test (%)**
Baseline (MCT)	-	-	-	-	30.55	24.31
DAN	√	√	√	√	25.13	21.15
① CRN	×	×	-	√	26.63	22.62
② CRN+AEP	×	√	√	√	27.12	23.22
③ CRN+AGP	√	×	√	√	25.58	21.45
④ DAN w/o *f*-MSE	√	√	×	√	30.36	25.33
⑤ DAN w/o DA	√	√	√	×	26.16	22.36
⑥ CRN w/o DA	×	×	-	×	26.68	22.95

waveform to log-fbank, WERs are reduced by SERGAN-fbank under low SNR levels ($\leqslant$0dB). SERGAN-fbank, MaskCoGAN and DereverbGAN have the similar model architecture and adversarial enhancement process as DAN, but all of them lack the proposed adversarial generation process and functional mean square error loss. As a result, our DAN achieves lower WERs under all evaluated SNR levels on the test set than the compared GAN-based methods. Specifically, the proposed DAN achieves 13.00% average relative WER improvements over the MCT ASR system on the test set. In addition, while other GAN-based methods degrade the recognition performance of the MCT ASR system under high SNR levels (>0 dB), our method improve the performance under all evaluated SNR levels (except 9 dB). This indicates that our DAN has a better noise robustness than all of the compared GAN-based methods.

4.2. Ablation study

The results of ablation studies are shown in Table 3. We first remove all adversarial processes from DAN, resulting in the fully supervised model, CRN. By comparing experiment ① with DAN, we can see that the WERs on both development and test sets increase significantly without the adversarial processes, which indicates that the proposed double adversarial training strategy improves the performance on ASR task. Then, we evaluate the effect of adversarial processes one by one. Experiment ② shows that simply performing the adversarial training between the enhancement model and discriminator even degrades the recognition performance in terms of WER. On the contrary, adding the adversarial generation process between the generator and discriminator improves the recognition performance. Experiments ② and ③ indicate that learning the speech distribution is crucial for adversarial training based enhancement models, which is missed in previous GAN-based methods. By comparing experiment ③ and DAN, we find that learning the differences between the clean and enhanced speeches can further improve the recognition performance. Experiment ④ indicates that the proposed f-MSE can efficiently utilize the representations learned by the discriminator, which is crucial for DANs in terms of WER. By comparing experiment ⑤ and ⑥, we can see that, without data augmentation, the proposed DAN still obtains lower WER than the fully supervised method. On the contrary, when the data augmentation is employed, DAN achieves another 1.21% WER reduction, which is better than the supervised method, CRN (0.33% WER reduction). This may indicate that DANs can utilize the data more efficiently.

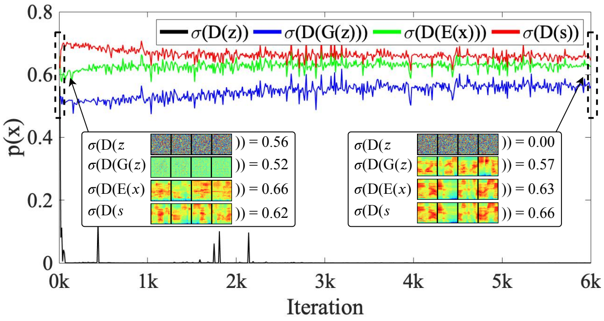

Figure 2: *The outputs of discriminator for random noises $\boldsymbol{z}$, generated speeches $\mathbf{G}(\boldsymbol{z})$, enhanced speeches $\mathbf{E}(\boldsymbol{x})$ and clean speeches $\boldsymbol{s}$ through the training process. For clarity, the outputs are activated by the sigmoid function σ.*

4.3. Exploring the outputs of models

Figure 2 plots the outputs of $\mathbf{D}$, $\mathbf{G}$ and $\mathbf{E}$ through the training process. We can see that the adversarial training process of DAN is very stable, which is import for adversarial training based methods. In the early stage of training, the discriminator cannot distinguish the enhanced and clean speeches correctly. Meanwhile, the generator only produces the noise-like samples, which are not similar to the clean speech yet. Through the double adversarial training, the discriminator becomes a powerful classifier with meaningful representations, which can distinguish the random, generated, enhanced and clean samples. Thanks to the discriminator, the generator produces speech-like samples, and the enhanced speeches are very similar to their real clean counterparts, resulting in the WER reduction.

5. Conclusions

In this paper, we propose the DAN-based monaural speech enhancement method for robust ASR, which consists of a discriminator, a generator and an enhancement model. By performing the adversarial training between the generator and discriminator, the representations of clean speeches are learned by the discriminator. Meanwhile, through the second adversarial process of the enhancement model and discriminator, the learned information is propagated to the enhancement model to guide its training. According to the experimental results on CHiME-2, we find that the proposed DAN significantly outperforms four recent GAN-based methods in terms of WER. Furthermore, ablation studies show that learning the speech distribution and using the proposed f-MSE are crucial for the robustness of speech recognition, which are missed in previous methods. In addition, the training process of DAN is very stable, which is important for adversarial training based methods.

6. References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning : An overview," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *INTERSPEECH*, 2013, pp. 436–440.

[3] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *ISMIR*, 2018, pp. 334–340.

[4] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *INTERSPEECH*, 2018, pp. 3229–3233.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.

[6] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *INTERSPEECH*, 2017, pp. 3642–3646.

[7] A. Pandey and D. Wang, "On adversarial training and loss functions for speech enhancement," in *ICASSP*, 2018, pp. 5414–5418.

[8] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[9] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001.

[10] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," in *ICLR*, 2019.

[11] D. Baby and S. Verhulst, "SERGAN : Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *ICASSP*, 2019, pp. 106–110.

[12] K. Wang, J. Zhang, S. Sun, Y. Wang, F. Xiang, and L. Xie, "Investigating generative adversarial networks based speech dereverberation for robust speech recognition," in *INTERSPEECH*, 2018, pp. 1581–1585.

[13] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *ICASSP*, 2018, pp. 5024–5028.

[14] B. Liu, S. Nie, Y. Zhang, D. Ke, S. Liang, and W. Liu, "Boosting noise robustness of acoustic model via deep adversarial training," in *ICASSP*, 2018, pp. 5034–5038.

[15] Z. Meng, J. Li, Y. Gong, and B.-h. F. Juang, "Adversarial feature-mapping for speech enhancement," in *INTERSPEECH*, 2018, pp. 3259–3263.

[16] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *ICLR*, 2016.

[17] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *CVPR*, 2017.

[18] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein GANs," in *NeurIPS*, 2017, pp. 5768–5778.

[19] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *ICLR*, 2018.

[20] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'chime' speech separation and recognition challenge: datasets, tasks and baselines." in *ICASSP*, 2013.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *ASRU*, 2011.

[22] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *ICASSP*, 2009, pp. 3761–3764.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[24] D. A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *ICLR*, 2016.

[25] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Sepration*, 2014, pp. 349–368.

[26] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[27] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *ICASSP*, 2019, pp. 6865–6869.