Adversarial autoencoder for reducing nonlinear distortion

Naohiro Tawara*, Tetsunori Kobayashi*, Masaru Fujieda†, Kazuhiro Katagiri†, Takashi Yazu†, and Tetsuji Ogawa*
* Waseda University, Tokyo, Japan
† OKI Electric Industry Co., Ltd., Saitama, Japan

***Abstract*—A novel post-filtering method using generative adversarial networks (GANs) is proposed to correct the effect of a nonlinear distortion caused by time-frequency (TF) masking. TF masking is a powerful framework for attenuating interfering sounds, but it can yield an unpleasant distortion of speech (e.g., a musical noise). A GAN-based autoencoder was recently shown to be effective for single-channel speech enhancement, however, using this technique for the post-processing of TF masking cannot help in nonlinear distortion reduction because some TF components are missing after TF-masking. Furthermore, the missing information is difficult embed using an autoencoder. In order to recover such missing components, an auxiliary reference signal that includes the target source components is concatenated with an enhanced signal, is then used as the input to the GAN-based autoencoder. Experimental comparisons show that the proposed post-filtering yields improvements in speech quality over TF-masking.**

I. Introduction

Speech enhancement aims at eliminating unexpected harmful noise from microphone observations in order to improve the quality and intelligibility of speech, which is important in wave generation for telecommunications and hearing aids. In addition, reducing noise plays an important preprocessing role in automatic speech recognition (ASR) in noisy environments.

Existing speech enhancement techniques can be categorized as linear or nonlinear approaches. TF masking [1] is a typical example of the latter approach, which attenuates adverse components, such as the interference source and diffuse noise, using a nonlinear filter that only passes TF components of the target source. Such nonlinear processing generally performs well in reducing interfering components, but it tends to unduly delete target source components as well, inducing unpleasant distortions, referred to as musical noise. Temporal smoothing [2] in a cepstral domain is able to remove such nonlinear distortion, but it induces other reverberation-like distortions. Denoising autoencoders (DAEs) are often used in speech enhancement problems and have been shown to outperform conventional nonlinear methods [3]. While DAEs provide certain improvements in applications, it is known that the mean squared errors used in their optimization cause over-smoothing and clip off speech segments [4], [5], [6]. To address this problem, adversarial structures [7] are incorporated into DAEs to constrain the neural network and, thus, generate realistic (i.e., unsmoothed) signals [8], [9], [10]. Pascual *et al.* proposed a GAN-based end-to-end speech enhancement method called a speech enhancement GAN (SEGAN), which generates signals in waveforms [8]. Donahue *et al.* extended the SEGAN model from a time domain to a TF domain, and demonstrated its effectiveness using ASR experiments [10].

Inspired by these works, an attempt is made to introduce an adversarial DAE as a post-filter of TF-masking to suppress nonlinear distortions contained in enhanced signals. Note that the missing components of the target source in the nonlinear distortions play an dominant role in decreasing speech intelligibility and ASR performance. However, it is difficult to restore these missing components from enhanced signals, even with the adversarial DAE and without any auxiliary information on the missing components. In addition, it is assumed that the distortion caused by nonlinear processing depends significantly on the type of interfering noise. Thus, auxiliary information on the target source and the interfering noise are introduced in order to train the adversarial DAEs (e.g., noise-aware training). Specifically, the estimated noise and observed signal are used as an auxiliary reference signal, and then as the input to a SEGAN, together with to the original noise-corrupted input. Exploiting auxiliary reference signals could be useful in improving the quality of enhanced speech affected by nonlinear distortions.

The two main contributions of the present work are: 1) showing the effectiveness of incorporating auxiliary information on the target source and the noise into a SEGAN, and 2) showing the effectiveness of a SEGAN in attenuating the nonlinear distortion of enhanced signals resulting from TF-masking.

The rest of the paper is organized as follows. Section II briefly explains conventional GAN-based speech enhancement. Section III describes the proposed GAN-based post filter of the nonlinear speech enhancement system. Section IV demonstrates the effectiveness of the proposed system by means of experiments on multichannel speech signals with an interference source. Finally, Section V concludes the paper.

II. Speech enhancement with GAN (SEGAN)

SEGAN [8] is a type of DAEs that gives a mapping from a noise-corrupted signal to a denoised signal. By incorporating an adversarial structure into the DAE, a SEGAN successfully generates realistic denoised signals that are difficult to distinguish from actual clean signals.

A SEGAN is composed of two networks: a generator G and a discriminator D. The generator is a denoising autoencoder

composed of 11 one-dimensional convolutions of 1×31 filters with a stride of two down-samples. The encoder receives a one-second waveform (i.e., 16384 samples at 16 kHz), applies 11 convolutions, and increases the depth of the filter, layer by layer. The result is an eight-dimensional feature map at the bottle-neck with a depth of 1024. The time-length $\times$ depth of the outputs of the layers are $16384 \times 1, 8192 \times 16, 4096 \times 32, 2048 \times 32, 1024 \times 64, 512 \times 64, 256 \times 128, 128 \times 128, 64 \times 256, 32 \times 256, 16 \times 512$, and 8×1024, respectively. Here, a noise vector is concatenated with the output of the encoder. The obtained latent vector is input to an up-sampling decoder, composed of 11 one-dimensional deconvolutions that have the same size filters and strides as those of the encoder. The output of each deconvolution layer is concatenated with the output of the homologous layer in the encoder. These skipped connections contribute to passing on fine-grained, low-level information from the encoder to the decoding stage [11], making optimization easier [12]. The activation function used in the encoder and decoder is a parametric rectified linear unit (PReLU) [13]. The $L1$ loss between clean and denoised signals is used to train the training DAEs.

The conditional discriminator extracts a feature-map from the denoised signals obtained from the encoder to a clean signal with convolution layers. The configurations of these convolutions are the same as those of the encoder, except that the activation function is a leaky ReLU instead of a PReLU. In addition, virtual batch normalization [14] is applied after each deconvolution to make the optimization faster. The 8×1024 dimensional feature map obtained after 11 convolution layers is converted into an 8×1 vector by a 1×1 convolution, and then aggregated into a single decision by a fully connected layer.

In training phase, encoder and decoder are alternately optimized with following adversarial procedure. First, fixing the parameter of generator G, the parameter of discriminator D is optimized by minimizing following loss function:

$$\mathcal{L}_{\mathrm{cGAN}}(D) = \mathbb{E}_{(\mathbf{x},\mathbf{x}_c)\sim p_{\mathrm{data}}(\mathbf{x},\mathbf{x}_c)}[(1 - D(\mathbf{x},\mathbf{x}_c))^2] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\mathbf{x}\sim p_{\mathrm{data}}(\mathbf{x})}[(D(\mathbf{x}, G(\mathbf{z},\mathbf{x})))^2] \quad (1)$$

where $p_{\mathrm{data}}(\mathbf{x}_c)$ denotes an empirical distribution over clean signal $\mathbf{x}_c$, $p_{\mathrm{data}}(\mathbf{x},\mathbf{x}_c)$ denotes an empirical distribution over a pair of observed signal $\mathbf{x}$ and clean signal $\mathbf{x}_c$, and $p_{\mathbf{z}}(\mathbf{z})$ denotes Gaussian distribution over a noise vector $\mathbf{z}$. By minimizing eq. (1), discriminator D try to discriminate whether the input is clean or denoised signal. Then, fixing the parameters of discriminator, the generator G is optimized by minimizing following loss function:

$$\mathcal{L}_{\mathrm{cGAN}}(G) = \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),(\mathbf{x},\mathbf{x}_c)\sim p_{\mathrm{data}}(\mathbf{x},\mathbf{x}_c)}[1 - D(\mathbf{x}, G(\mathbf{z},\mathbf{x}_c)))^2 + \lambda||\mathbf{x} - G(\mathbf{z},\mathbf{x}_c)||_1] \quad (2)$$

where λ is a weight between adversarial and reconstruction losses. We set it to 100 for the training. By minimizing eq. (1), the generator G try to generated denoised signals which is difficult to distinguish from clean signals. After alternating optimization of eqs. 1 and 2, Denoised signals of high quality are obtained from the generator G.

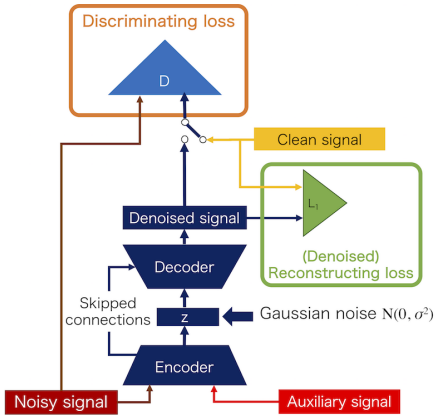


Fig. 1: Architecture of a speech enhancement generative adversarial network (SEGAN) with an auxiliary reference input.

III. Nonlinear distortion compensation using a SEGAN with auxiliary inputs

Enhanced speech with a nonlinear distortion is extracted from microphone observations using TF masking, and then input to a nonlinear distortion compensation system based on a SEGAN with auxiliary inputs. Note that a mapping from enhanced speech to clean speech is difficult to learn because target source components may have been deleted from the enhanced speech as a result of the nonlinear processing. In addition, the characteristics of the distortion are closely related to the type of interfering noise. Thus, noise information could be effective in recovering these missing components. To achieve this using a SEGAN, auxiliary reference signals on the target source and the interfering noise are used as inputs to the encoder, in addition to the enhanced signals. Figure 1 shows the overall structure of the proposed SEGAN with auxiliary reference inputs.

IV. Speech enhancement experiment

Experiments were used to compare the sound source separation in order to demonstrate the effectiveness of exploiting auxiliary information in SEGAN-based nonlinear distortion compensation.

A. Experimental setup

1) Speech material: Figure 2 shows the experimental environment. The target source was placed in front of two microphones, and the interference source was placed next to the microphones (i.e., 90 degrees to the target). In this experiment, sound source segregation based on estimating incident angle of each Frequency component of Input signals acquired by multiple microphones (SAFIA), configured to enhance the front area of microphones using TF masking, was applied in the speech enhancement stage. Nine types

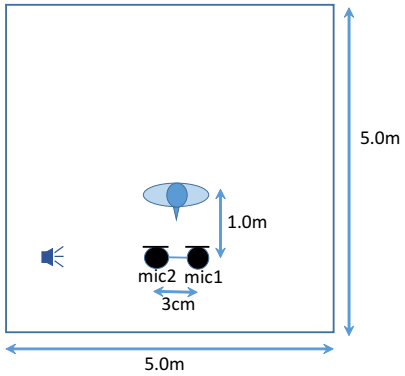


Fig. 2: Experimental environment with two microphones, a target source, and an interference source.

TABLE I: Interfering noise recorded. Noise signals are selected from the JEIDA noise database.

DB id	noise type	use
09	exhibition hall (booth)	training
11	exhibition hall (aisle)	training
13	station (concourse)	training
14	station (aisle)	training
18	factory (machine)	training
20	factory (metal)	training
26	street	training
28	intersection	training
30	crowd	testing
47	elevator hall	testing

of noise were chosen from the JEIDA noise corpus [15], and the impulse responses between the sound source and the microphones were recorded. Table I lists the types of noise samples used for training and testing.

The noise-corrupted signals were synthesized by convoluting the dry sources of speech with the impulse responses, and then mixing the convolved speech with the noise at five signal-to-noise ratios (SNRs) of -10, -5, 0, 5, and 10 dB. The dry sources were 8000 utterances spoken by 78 females selected from the JNAS [16], yielding about 50 different sentences for each speaker and noise condition. To create a test set, 100 convolved spoken utterances were corrupted with two unseen types of noise at five SNRs of -10, -5, 0, 5, and 10 dB. Note that the combinations of experimental conditions in terms of speakers, utterances, and noise types differed between training and testing.

2) Speech enhancement: In the present experiment, a phase-based SAFIA [17] was used in a nonlinear speech enhancement system. Speech-dominant signals were obtained by masking TF bins, where the difference between the phases of two channels was higher than 0.1, while the noise-dominant signals were extracted by masking the remaining bins.

3) Evaluation criteria: A signal distortion rate (SDR) between the estimated and the desired clean speech is calculated using the BSS Eval toolbox [18] to evaluate how well the nonlinear distortion is compensated. In order to measure the perceptual performance, a perceptual evaluation of speech quality (PESQ), based on the ITU standard P.862 [19], is also measured.

TABLE II: Models evaluated.

system	original input	auxiliary reference input
observation	noisy speech	——
SAFIA	noisy speech	——
SEGAN	noisy speech	——
SEGAN-oracle	noisy speech	matched (correct) noise
SEGAN-matched	noisy speech	matched (unsynchronized) noise
SEGAN-enhanced	noisy speech	enhanced noise (by SAFIA)
SAFIA-SEGAN	enhanced voice	——
SAFIA-SEGAN-oracle	enhanced voice	matched (correct) noise
SAFIA-SEGAN-matched	enhanced voice	matched (unsynchronized) noise
SAFIA-SEGAN-enhanced	enhanced voice	enhanced noise (by SAFIA)
SAFIA-SEGAN-obs	enhanced voice	microphone observation

B. Experimental results

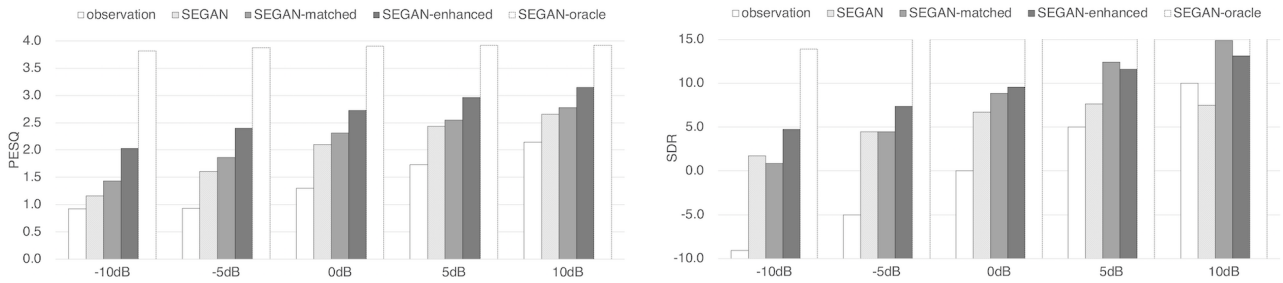
1) Effectiveness of using auxiliary information on the types of noise: The effectiveness of using auxiliary information on the types of noise as inputs to SEGAN was evaluated. In this case, the following four models were compared:

- **SEGAN:** original SEGAN, without any reference signals;
- **SEGAN-oracle:** SEGAN using an oracle noise signal, which is consistent with the convoluted signal from the dry source, as the auxiliary input.
- **SEGAN-matched:** SEGAN using the matched noise, which consists of non-speech segments extracted from the oracle noise (i.e., the alignment is not correct), as the auxiliary input.
- **SEGAN-enhanced:** SEGAN using the noise-dominant signal obtained by SAFIA as an auxiliary reference input.

Figure 3 shows the speech enhancement performance obtained by SEGANs with and without reference signals. The figure shows that the SEGAN yielded a notable improvement over the original noise-corrupted observation. In addition, further improvements were obtained by introducing any kind of reference signals. In particular, the best performance was obtained when using an oracle noise signal as a reference (**SEGAN-oracle**). This result indicates that a SEGAN could learn a specific filter that calculated the differences between observed and noise signals when oracle noise signals were provided. However, the oracle noise is not available in practice. Instead, we obtain similar effects by using matched or enhanced noises (**SEGAN-matched** and **SEGAN-enhanced**, respectively).

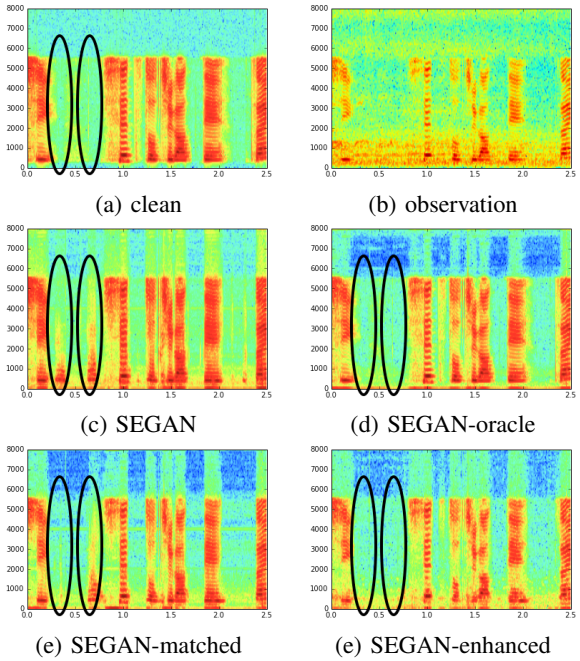
Figure 4 depicts the spectrograms of the signals obtained from each model. The figure shows that all signals except SEGAN-enhanced generated redundant components around the regions marked by circles. Thus, SEGAN-enhanced achieves the best performance of the proposed models.

2) Effectiveness of SEGAN for nonlinear distortion compensation: The effectiveness of SEGAN for nonlinear distortion compensation was evaluated. In this case, the following five models were compared:



(a) PESQ (b) SDR

Fig. 3: Speech enhancement performance of SEGANs with and without auxiliary reference inputs, where PESQ and SDR were averaged over 10 utterances for each condition.



(a) clean (b) observation

(c) SEGAN (d) SEGAN-oracle

(e) SEGAN-matched (e) SEGAN-enhanced

Fig. 4: Spectrograms of (a) a clean signal and (b) an observed noise-corrupted signal, and enhanced signals obtained by (c) an original SEGAN, (d) SEGAN-oracle, (e) SEGAN-matched, and (f) SEGAN-enhanced.

- **SAFIA-SEGAN:** original SEGAN applied on enhanced signals obtained from SAFIA, without any reference signals;
- **SAFIA-SEGAN-oracle:** SAFIA-SEGAN using an oracle noise signal as an auxiliary reference input.
- **SAFIA-SEGAN-matched:** SAFIA-SEGAN using the matched noise as an auxiliary reference input.
- **SAFIA-SEGAN-enhanced:** SAFIA-SEGAN using the noise-dominant signal as an auxiliary reference input.
- **SAFIA-SEGAN-obs:** SAFIA-SEGAN using the original observation as an auxiliary reference input.

Figure 5 shows the speech enhancement performance obtained by SAFIA-SEGANs with and without reference signals. This figure shows that **SAFIA-SEGAN** yielded a notable improvement over **SAFIA**. This result indicates that SEGAN is effective in compensating for nonlinear distortion. Further improvements were obtained by introducing an enhanced signal as a reference. On the other hand, PESQ and SDR deteriorated when a matched signal was introduced as a reference (**SAFIA-SEGAN-matched**). Note that SDR deteriorated, while PESQ improved when an observation was introduced as a reference (**SAFIA-SEGAN-obs**). This is because the removed components could be recovered, but faint noise signals were added to the output.

Figure 6 shows the spectrograms of the signals obtained from each model. From this figure, we can see that SAFIA removed too many components, generating the musical noise. Furthermore, SAFIA-SEGAN contributed to attenuating the nonlinear distortion, and introducing enhanced or observed signals provided further improvements.

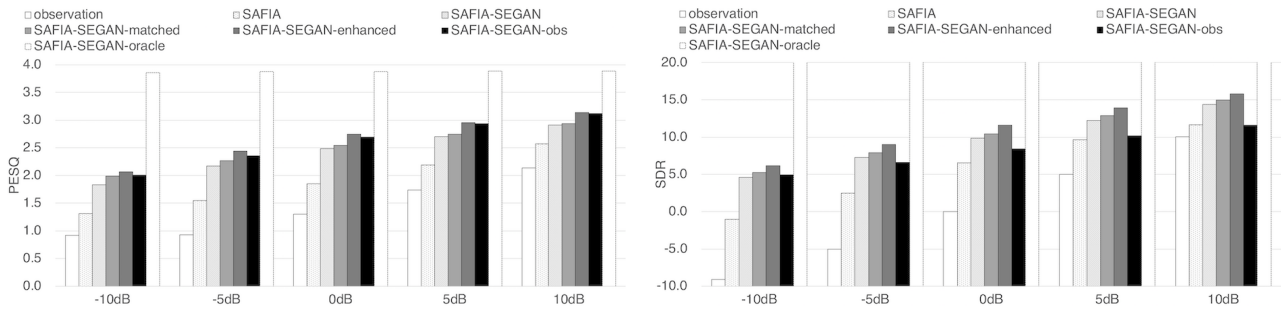
V. Conclusions

This study proposed a novel post filtering method using a GAN to correct the nonlinear distortion caused by TF masking. We showed that simply using a GAN on the output of TF masking cannot reduce nonlinear distortion because some TF components are missing after TF-masking. In order to solve this problem, an estimated noise signal was concatenated with an enhanced signal, and then used as the input to a GAN-based autoencoder. Experiment results showed that the proposed post-filtering method yielded improvements in speech quality over TF masking.

References

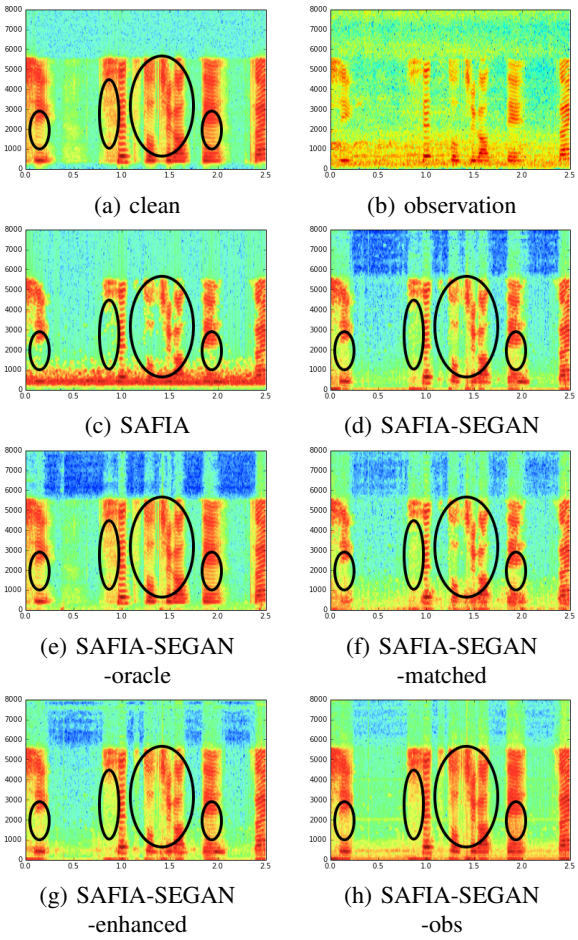
[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on signal processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[2] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *ICASSP*. IEEE, 2008, pp. 45–48.



(a) PESQ (b) SDR

Fig. 5: Speech enhancement performance from SAFIA and SAFIA-SEGAN, with and without auxiliary reference signals, where PESQ and SDR were averaged over 10 utterances for each condition.



(a) clean (b) observation

(c) SAFIA (d) SAFIA-SEGAN

(e) SAFIA-SEGAN -oracle (f) SAFIA-SEGAN -matched

(g) SAFIA-SEGAN -enhanced (h) SAFIA-SEGAN -obs

Fig. 6: Spectrogram of (a) a clean signal and (b) an observed noise-corrupted signal, and enhanced signal obtained by (c) SAFIA, (d) SAFIA-SEGAN, (e) SAFIA-SEGAN-oracle, (f) SAFIA-SEGAN-matched, (g) SAFIA-SEGAN-enhanced, and (h) SAFIA-SEGAN-obs, respectively. Note that band-pass filter with band 300–5500 Hz was applied for TF-masking.

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

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

[5] P. G. Shivakumar and P. G. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement." in *INTERSPEECH*, 2016, pp. 3743–3747.

[6] T. G. Kang, J. W. Shin, and N. S. Kim, "Dnn-based monaural speech enhancement with temporal and spectral variations equalization," *Digital Signal Processing*, vol. 74, pp. 102–110, 2018.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in NIPS*, 2014, pp. 2672–2680.

[8] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[9] S. Pascual, M. Park, J. Serrà, A. Bonafonte, and K.-H. Ahn, "Language and noise transfer in speech enhancement generative adversarial network," *arXiv preprint arXiv:1712.06340*, 2017.

[10] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," *arXiv preprint arXiv:1711.05747*, 2017.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[13] "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[14] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[15] S. Itahashi, "A noise database and japanese common speech data corpus," *The Journal of the Acoustical Society of Japan*, vol. 47, no. 12, pp. 951–953, 1991.

[16] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[17] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoustical Science and Technology*, vol. 22, no. 2, pp. 149–157, 2001.

[18] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[19] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.