Noise Prior Knowledge Learning for Speech Enhancement via Gated Convolutional Generative Adversarial Network

Cunhang Fan*† Bin Liu* Jianhua Tao*†‡ Jiangyan Yi* Zhengqi Wen* and Ye Bai*†
* National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
† School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
‡ CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
E-mail:{cunhang.fan, liubin, jhtao, jiangyan.yi, zqwen, ye.bai}nlpr.ia.ac.cn

***Abstract*—Speech enhancement generative adversarial network (SEGAN) is an end-to-end deep learning architecture, which only uses the clean speech as the training targets. However, when the signal-to-noise ratio (SNR) is very low, predicting clean speech signals could be very difficult as the speech is dominated by the noise. In order to address this problem, in this paper, we propose a gated convolutional neural network (CNN) SEGAN (GSEGAN) with noise prior knowledge learning to address this problem. The proposed model not only estimates the clean speech, but also learns the noise prior knowledge to assist the speech enhancement. In addition, gated CNN has an excellent potential for capturing long-term temporal dependencies than regular CNN. Motivated by this, we use a gated CNN architecture to acquire more detailed information at waveform level instead of regular CNN. We evaluate the proposed method GSEGAN on Voice Bank corpus. Experimental results show that the proposed method GSEGAN outperforms the SEGAN baseline, with a relative improvement of 0.7%, 28.2% and 43.9% for perceptual evaluation of speech quality (PESQ), overall Signal-to-Noise Ratio (SNRovl) and Segmental Signal-to-Noise Ratio (SNRseg), respectively.**

I. Introduction

The goal of speech enhancement is to remove the noise from an observed signal recorded in noisy environment. It has been widely used in many applications, such as automatic speech recognition (ASR) [1], speech coding [2] and hearing aids [3].

In the past, various speech enhancement methods have been developed [4], [5], [6], [7], [8]. Notable examples include spectral subtraction [4], Wiener filtering [5] and nonnegative matrix factorization (NMF) [6]. In recent years, deep learning has achieved state-of-the-art performance in many applications. Motivated by the success of deep learning, researchers have developed many deep learning techniques for speech enhancement, such as deep denoising auto-encoders [7], deep neural networks (DNNs) [8], and convolutional neural networks (CNNs) [9]. These DNN-based speech enhancement models learn a mapping between noisy input features and the desired target signals. However, when the signal-to-noise ratio (SNR) is very low, predicting clean speech features could be very difficult as the speech is dominated by the noise. In order to address this issue, Odelowo et al. [10] propose a noise prediction method for speech enhancement. They use the noise as their target features instead of clean speech. However, they only predict the noise signals and don't make full use of the clean speech, which may lead to speech distortion.

In convolutional neural networks (CNNs), when the receptive fields are expanded, contextual information can be augmented. In order to achieve this goal, there are two main methods. One way is to increase the network depth, which decreases computational efficiency and typically results in vanishing gradients [11]. Another way is to enlarge the kernel size, but it will raise computational burden and training time. Recent works [12], [11], [13] have shown that CNNs with gating mechanisms have an excellent potential for capturing long-term temporal dependencies. Li et al. [12] apply the gated CNN for speech separation and the performance is improved. Tan et al. [11] use the gated CNN for speech enhancement and get a good performance. However, they only enhance the magnitude spectrum of complex-valued short time Fourier transform (STFT) coefficients, leaving the phase spectrum unchanged. Recent studies [14], [15] show that the perceptual quality of estimated signals can be improved by enhancing the phase spectrum.

In order to make full use of the raw data, Pascual et al. [16] do speech enhancement based on raw waveform directly via generative adversarial network (GAN). Their generator network is structured similarly to an auto-encoder via CNN. But they only enhance the target signals, the noise information is not utilized. In this paper, motivated by the success of gated CNN [11], [12], [13] and speech enhancement GAN (SEGAN) [16], we propose a gated convolutional GAN method with noise prior knowledge for speech enhancement, named as GSEGAN. Because the gated CNN has advantages in capturing long-term dependencies in sequential data, all the network architectures of the proposed system are built with gated CNN rather than the regular CNN. In addition, many traditional methods usually use noise estimation for speech enhancement, for example Wiener filtering [5]. However, DNN-based methods pay insufficient attention to this. When the SNR is very low, predicting clean speech can be very difficult as the noise dominates the speech signals. To address this problem, our proposed method not only estimates the clean signals from

noisy speech, but also makes full use of background-noise. Noise signals are learned from noisy input and the enhanced speech. Then the noise prior knowledge is used to instruct the noise signals learning and assist the speech enhancement. In this way, the knowledge of estimated noise can be used by enhanced signals to improve the performance of speech enhancement.

The rest of this paper is organized as follows. In section 2, speech enhancement generative adversarial network is presented. The proposed method for speech enhancement is stated in section 3. Section 4 shows detailed experiments and results. Section 5 draws conclusions.

II. Speech Enhancement Generative Adversarial Network

GAN is firstly introduced by Goodfellow et al. in [17], which consists of a generator G and a discriminator D. The generator G maps a noise vector z, from some known prior distribution $p_z(z)$, to fake samples $G(z)$. The main task of discriminator D is to recognize whether its input is from training data (real) or G (fake). G and D are pitted against each other in an adversarial framework.

Speech enhancement generative adversarial network (SEGAN) [16] applies the GAN to speech enhancement. Generator (G) network is structured similarly to an auto-encoder via CNN. The G network performs the enhancement. They employ the least-squares GAN (LSGAN) [18] to their speech enhancement system. The G loss is defined as follow:

$$\min_G J(G) = \frac{1}{2}\mathbf{E}[(D(G(\boldsymbol{x},\boldsymbol{z}))-1)^2] + \lambda||G(\boldsymbol{x},\boldsymbol{z})-\boldsymbol{y}||_1 \quad (1)$$

where $\boldsymbol{z}$ denotes the noise sample from normal distribution $\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$. $\boldsymbol{x}$ and $\mathbf{y}$ are noisy input and the target speech, respectively.

III. The Proposed Speech Enhancement System

The main task of speech enhancement is to obtain the enhanced signals $\widetilde{\boldsymbol{y}}$ from the noisy input signals $\boldsymbol{x}$. In this paper, we propose to do so with a gated CNN based speech enhancement GAN (GSEGAN) system. Moreover, the proposed system not only estimates the clean signals from noisy speech, but also makes full use of background-noise to improve the performance of speech enhancement. In other words, the noise prior knowledge is learned to help to obtain better enhanced signals. The gated CNN has an excellent potential for capturing long-term temporal dependencies so that it can deal with the raw waveform better than regular CNN. Motivated by the recent success achieved by gated CNN in speech enhancement [11], [12], in this paper, we propose to use gated CNN for all the convolutional network architectures.

A. Gated CNN

Gating mechanisms potentially facilitate modeling more complex interactions by controlling the information flow [11], which is shown in the Figure 1. The proposed methods use the following gated activation unit:

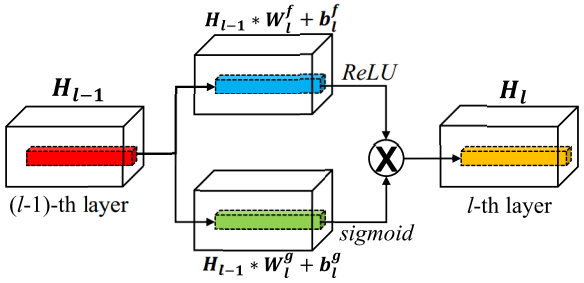

Fig. 1. The architecture of gated CNN.

$$H_l = ReLU(H_{l-1} * W_l^f + b_l^f) \otimes \sigma(H_{l-1} * W_l^g + b_l^g) \quad (2)$$

where H_{l-1} and H_l denote the output of the (l-1)-th and l-th layer. W_l^f, W_l^g, b_l^f and b_l^g represent kernels and biases of the l-th layer. ReLU, σ, $*$ and $\otimes$ are ReLU activation, sigmoid activation, convolution operation and the element-wise multiplication, respectively.

B. Network architectures

In this study, we propose to enhance the noisy speech by the deep adversarial training method, named GSEGAN. The model consists of a generator (G) and a discriminator (D). The G network performs the enhancement, which is shown in the Figure 2. It transforms the noisy speech into the enhanced signals. The main task of discriminator D is to distinguish between the enhanced signals and clean ones.

The G network is an encoder-decoder, adapted from [16]. It consists of symmetric encoding layers and decoding layers. In the encoder stage, the input noisy $\boldsymbol{x}$ is a linearly mixed single.

$$\boldsymbol{x} = \boldsymbol{y} + \boldsymbol{n} \quad (3)$$

where $\boldsymbol{y}$ and $\boldsymbol{n}$ are clean signal and the noise, respectively. Then $\boldsymbol{x}$ is protected and compressed through a number of strided convolutional layers.

$$\boldsymbol{c} = f_e(\boldsymbol{x}) \quad (4)$$

where vector $\boldsymbol{c}$ is a condensed representation, $f_e(*)$ is a mapping function from the input features to the vector $\boldsymbol{c}$.

In the decoding stage, the encoding process is reversed by means of fractional strided transposed convolutions (sometimes called deconvolutions).

$$\widetilde{\boldsymbol{y}} = f_d(\boldsymbol{c},\boldsymbol{z}) \quad (5)$$

where $\widetilde{\boldsymbol{y}}$ is the enhanced signal, $f_d(*)$ is a mapping function of the decoder. z denotes a latent vector, which is the noise sample from normal distribution $\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$.

Skip connections are also used in the G network, which connect each encoding layer to its corresponding homologous decoding layer. The convolutional feature maps are passed to and summed with the deconvolutional feature maps element-wise, and passed to the next layer (Figure 2). If we force all the information to flow through the compression bottleneck,

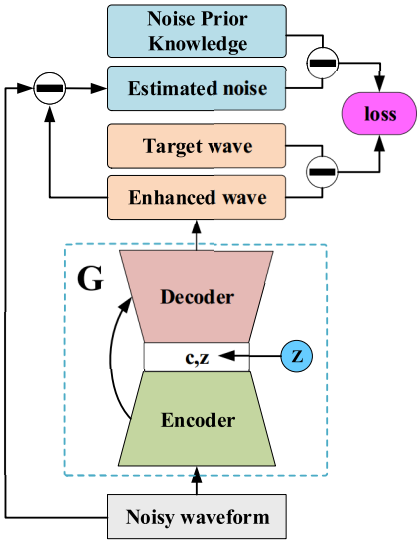

Fig. 2. The architecture of the proposed adversarial training method. **c** is a condensed representation after the encoder stage. **z** denotes a latent vector, which is the noise sample from normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

many low level details could be lost, but with skip connections the speech waveform can be reconstructed properly. Moreover, skip connections can offer a better training behavior, as the gradients can flow deeper through the whole structure without suffering much vanishing [19].

C. Noise prior knowledge learning

When the noise dominates the speech signals, predicting clean speech can be very difficult. Therefore, only using the clean speech as the training targets may be insufficient for speech enhancement. In order to improve the performance of speech enhancement, we use the noise prior knowledge to instruct the learning of noise and assist the predicting of target signals.

Because the proposed method operates the speech enhancement in the time domain, so the noise signal $\boldsymbol{n}$ can be estimated by the input noisy speech $\boldsymbol{x}$ and the enhanced speech $\widetilde{\boldsymbol{y}}$, which is motivated by [20]:

$$\widetilde{\boldsymbol{n}} = \boldsymbol{x} - \widetilde{\boldsymbol{y}} \tag{6}$$

where $\widetilde{\boldsymbol{n}}$ is the estimated noise. Then the noise prior knowledge is used to adjust the estimated noise.

D. Loss function

In order to improve the performance of speech enhancement, the noise prior knowledge is used as a regularization at the loss function, which is shown as following:

$$J_n = \| \boldsymbol{n} - \widetilde{\boldsymbol{n}} \|_1 \tag{7}$$

To measure the distance between the enhanced speech and the clean examples, we choose the L_1 norm, because it has been proven to be effective in the image manipulation domain [21], [22]. Therefore, the loss function in Eq.1 becomes:

$$\begin{aligned} \min_G J(G) = \frac{1}{2}\mathbf{E}[(D(G(\boldsymbol{x}, \boldsymbol{z})) - 1)^2] \\ + \lambda[\| \widetilde{\boldsymbol{y}} - \boldsymbol{y} \|_1 + \alpha J_n]. \end{aligned} \tag{8}$$

where λ and α are the weight of L_1 regularization and noise estimation, $\boldsymbol{y}$ is the target signals. When $\alpha = 0$, it means that it does not use the noise prior knowledge and the loss function of G is the same as SEGAN [16].

IV. Experiments and Results

A. Dataset

In order to compare with the SEGAN [16], we chose the same dataset Voice Bank corpus [23]. The database is open and available[1]. We select 30 speakers from this database for our experiments. The sampling rate is 16kHz.

As for the training set, a noise database is used [24], which includes 40 different conditions and 10 types of noises. In each condition, every speaker has about 10 different sentences mixed with the noise at 4 signal-to-noise ratio (SNR) (0, 5, 10 and 15 dB). The test set includes 20 different conditions and 5 types of noise with 4 SNR each (2.5, 7.5, 12.5 and 17.5 dB). For each test speaker, there are around 20 different sentences in each condition. Note that all test utterances are excluded from the training set, using different speakers and conditions.

B. Experimental setup

During training, we extract chunks of waveforms with a sliding window and the chunks are approximately one second of speech (16384 samples) with 50% overlap.

In the gated CNN, the ReLU and sigmoid layers have the same settings. In (in-Channel, out-Channel) format, the encoder stage in G network has (1, 16), (16, 32), (32, 32), (32, 64), (64, 64), (64, 128), (128, 128), (128, 256), (256, 256), (256, 512), (512, 1024) convolution layers with 31×1 kernels and 2×1 strides. The input size is 16384×1. As for the decoder stage of G, it is a mirroring of the encoder with the same settings.

The D is a convoluntional classification network, which follows the same convolutional structure of G's encoder stage. However, there are 3 differences. First, the input size is 16384×2. Second, it uses LeakyReLU [25] non-linearities with $\alpha = 0.3$. Finally, a flatten layer is added at the last, which reduces the amount of parameters required for the final classification neuron.

In all the experiments, the epoch is set to 86, batch size is 100 and learning rate is 0.0002. Our models are optimized with RMSprop algorithm [26] and implemented using Tensorflow deep learning framework. The λ weight of L_1 regularization is 100, which is same as SEGAN [16].

[1]http://dx.doi.org/10.7488/ds/1356

TABLE I
THE RESULTS OF PESQ, SNROVL AND SNRSEG FOR DIFFERENT SPEECH ENHANCEMENT METHODS. α IS THE WEIGHT OF THE NOISE PRIOR KNOWLEDGE. THE VALUES OF THOSE METRICS ARE THE HIGHER THE BETTER.

Method	α	$PESQ$	$SNRovl$	$SNRseg$
Noisy	-	1.970	8.446	1.680
SEGAN (baseline)	0	2.269	13.775	6.210
SEGAN +noise prior knowledge (proposed)	0.1	**2.319**	13.095	4.126
	0.5	2.275	15.922	7.680
	0.8	2.217	16.563	8.090
	1.0	2.224	14.813	6.558
GSEGAN (proposed)	0	2.205	15.447	6.737
	0.1	2.219	15.411	7.304
	0.5	2.226	17.165	8.611
	0.8	2.191	16.894	8.404
	1.0	2.284	**17.656**	**8.935**

C. Evaluation metric

In this work, in order to evaluate the quality of the enhanced speech, we compute the following objective measures (the higher the better). The perceptual evaluation of speech quality (PESQ) [27] measures. The overall Signal-to-Noise Ratio (SNRovl) and Segmental Signal-to-Noise Ratio (SNRseg) [28] are from 0 to ∞.

D. Results

Table I shows the results of these metrics. To have a comparative reference, it also shows the results of those metrics when applied directly to the noisy signals. We re-implement SEGAN [16] with our experiment setup and it is used as our baseline. It is corresponding to the SEGAN method and α=0 in Table I.

1) The effect of noise prior knowledge learning: Notice that, α=0 means that methods only predict the clean speech but with no noise prior knowledge. From the Table I we can find that the performance of the enhanced system can be improved when the noise prior knowledge is used. More specifically, when α=0.1, the proposed method based on SEGAN achieves 2.319 for PESQ. However, as for SEGAN baseline, it is only 2.269. Meanwhile, when α=0.8, the SNRovl and SNRseg of the proposed method based on SEGAN are 16.563 and 8.090. But as for SEGAN baseline, they are only 13.775 and 6.210. When the SNR is very low, predicting clean speech can be very difficult as noise dominates the speech signals. Therefore, when the noise prior knowledge is used, enhanced signals can borrow knowledge from the estimated noise. These results suggest that the noise prior knowledge can improve the performance of speech enhancement and reduce the distortion of target speech.

2) GSEGAN vs. SEGAN: Table I shows that when the regular CNN is replaced by gated CNN, GSEGAN (α=0) gets slightly worse PESQ than SEGAN. However, in all the other metrics (SNRovl and SNRseg), GSEGAN outperforms the SEGAN method. More specifically, the GSEGAN achieves 1.672 and 0.527 improvement for SNRovl and SNRseg compared with SEGAN. This indicates that the gated CNN can deal with the raw waveform better than regular CNN. Meanwhile, it also suggests that the advantage of gated CNN in dealing with the long-term temporal dependencies.

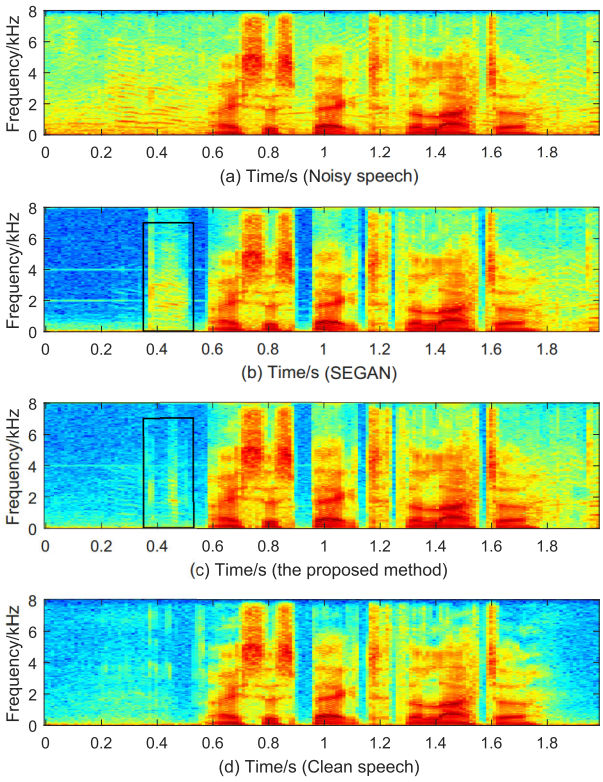

Fig. 3. An example of noisy, SEGAN, the proposed method and clean spectrogram for a speech segment from the test set. (a):spectrogram of the noisy speech; (b):SEGAN (baseline); (c):the proposed method ($\lambda = 1.0$); (d):clean speech.

In addition, when α gets larger, for example α=1.0, it means that all of the background-noise information is considered. Compared with SEGAN, the GSEGAN achieves a 0.06, 2.843 and 2.377 improvement for PESQ, SNRovl and SNRseg, respectively. It reveals the effectiveness of the GSEGAN in learning complex and detailed information at waveform level.

3) Evaluation of the proposed method : Finally, when the GSEGAN and noise prior knowledge are used simultaneously, the proposed methods are superior to the SEGAN baseline in most case. Especially, when α=1.0, the proposed method beats baseline SEGAN in all objective measures. More specifically, the proposed method gets 2.284 for PESQ, 17.656 for SNRovl and 8.935 for SNRseg. Compared with SEGAN, the proposed method produces a relative improvement of 0.7%, 28.2% and 43.9% for PESQ, SNRovl and SNRseg, respectively. These results confirm that the proposed GSEGAN with noise prior knowledge can improve the performance of the speech enhancement.

As an example, Figure 3 shows an example of noisy, SEGAN, the proposed method and clean spectrogram for a

speech segment from the test set. Notice that, compared with the spectrograms of clean speech, the harmonics of the proposed model are preserved well, and the formant structures are seen to be effectively preserved in the reconstructed speech. Those indicate that the noisy speech is effectively enhanced by the proposed model. On the other hand, the baseline SEGAN can enhance the noisy speech, but the formant structure is not clear compared with the proposed method. For example, compared with (c) in Figure 3, we can see that in (b) some formant structures are not reconstructed and some noise signals are not removed very well (marked in the black boxes).

V. Conclusions

In this work, we propose a gated convolutional GAN method with noise prior knowledge for speech enhancement. Different from SEGAN, the proposed method uses gated CNN instead of regular one. The reason is that the gated mechanisms have an excellent potential for capturing long-term structures. In order to address the low-SNR signals speech enhancement's problem, our proposed method not only predicts the clean speech, but also uses the noise prior knowledge to instruct the learning of noise and assist the estimating of target signals. Results show that the proposed method outperforms SEGAN baseline, with a relative improvement of 0.7%, 28.2% and 43.9% for PESQ, SNRovl and SNRseg, respectively. In the future, we will explore the proposed method for multi-channel speech enhancement.

VI. Acknowledgements

This work is supported by the National Key Research & Development Plan of China (No.2017YFB1002802), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61831022, No.61771472, No.61603390), the Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02050100), and Inria-CAS Joint Research Project (No.173211KYSB20170061).

References

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

[2] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with wiener filter for high-quality speech communication," *Speech Communication*, vol. 53, no. 5, pp. 677–689, 2011.

[3] H. Levitt, "Noise reduction in hearing aids: A review," *Journal of rehabilitation research and development*, vol. 38, no. 1, pp. 111–122, 2001.

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[5] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *Proc. ICASSP*, vol. 2. IEEE, 1996, pp. 629–632.

[6] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

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

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

[9] S. R. Park and W. L. Jin, "A fully convolutional neural network for speech enhancement," in *INTERSPEECH*, 2017, pp. 1993–1997.

[10] B. O. Odelowo and D. V. Anderson, "A study of training targets for deep neural network-based speech enhancement using noise prediction," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5409–5413.

[11] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for supervised speech separation," in *Proc. ICASSP*, 2018.

[12] L. Li and H. Kameoka, "Deep clustering with gated convolutional networks," in *Proc. ICASSP*, 2018, pp. 16–20.

[13] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *arXiv preprint arXiv:1612.08083*, 2016.

[14] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[15] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, 2015.

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

[17] 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 neural information processing systems*, 2014, pp. 2672–2680.

[18] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2813–2821.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *computer vision and pattern recognition*, pp. 770–778, 2016.

[20] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. ICASSP*. IEEE, 2018, pp. 5069–5073.

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[22] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

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

[24] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 146–152.

[25] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1, 2013, p. 3.

[26] T. Tieleman and G. Hinton, "Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning," Technical Report. Available online: https://zh. coursera. org/learn/neuralnetworks/lecture/YQHki/rmsprop-divide-the-gradient-by-a-running-average-of-its-recent-magnitude (accessed on 21 April 2017), Tech. Rep.

[27] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part i–time-delay compensation," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, 2002.

[28] S. Quackenbush, T. Barnwell, and M. Clements, "Objective measures of speech qualityprentice-hall," *Englewood Cliffs, NJ*, 1988.