

Generative Adversarial Network based Acoustic Echo Cancellation

Yi Zhang, Chengyun Deng, Shiqian Ma, Yongtao Sha, Hui Song, Xiangang Li

Didi Chuxing, Beijing, China

{yizhang, songhui}@didiglobal.com

Abstract

Generative adversarial networks (GANs) have become a popular research topic in speech enhancement like noise suppression. By training the noise suppression algorithm in an adversarial scenario, GAN based solutions often yield good performance. In this paper, a convolutional recurrent GAN architecture (CRGAN-EC) is proposed to address both linear and nonlinear echo scenarios. The proposed architecture is trained in frequency domain and predicts the time-frequency (TF) mask for the target speech. Several metric loss functions are deployed and their influence on echo cancellation performance is studied. Experimental results suggest that the proposed method outperforms the existing methods for unseen speakers in terms of echo return loss enhancement (ERLE) and perceptual evaluation of speech quality (PESQ). Moreover, multiple metric loss functions provide more freedom to achieve specific goals, e.g., more echo suppression or less distortion.

Index Terms: nonlinear echo cancellation, deep learning, generative adversarial network, convolutional recurrent network

1. Introduction

Acoustic echo originates in a local audio loop back that occurs when a (near-end) microphone picks up audio signals from a speaker, and sends it back to a (far-end) participant. Echo can be extremely disruptive to a conversation, and makes the phone calls very unpleasant and distracting. Acoustic echo cancellation (AEC) or suppression (AES) aims to suppress echo from microphone signal whilst leaving the speech of near-end talker least distorted. Conventional echo cancellation algorithms estimate the echo path by using adaptive filter [1], under the assumption of a linear relationship between far-end signal and acoustic echo. In practice this linear assumption does not always hold, and thus a post-filter [2] [3] is often deployed to suppress the residue echo. However, performance of such AEC algorithms drops dramatically when nonlinearity is introduced. Although some nonlinear adaptive filters like Volterra filter [4] were proposed, they are too expensive to implement.

With the advancement in deep learning, many of the speech processing tasks, including speech recognition [5], noise suppression [6] [7], speech separation [8] [9] have been done using deep neural networks. Several solutions are also proposed on acoustic echo cancellation. Lee et al. [10] used a deep neural network with 3 layers of restricted Boltzmann machine (RBM) to predict the gain of residual echo suppression. Muller et al. [11] suggested to use near-end inactive frequencies to adapt acoustic transfer function during double talk, where a two fully connection layer network is used to detect the activity of near-end signal. Zhang and Wang [12] proposed a bidirectional long-short term memory (BLSTM) to predict the ideal ratio mask

from microphone signals, which is then used to resynthesize the near-end speech. This solution does not need double talk detection whilst conventional methods do. Carbajal et al. [13] built a two-layer network to predict phase sensitive filter of the residual echo suppression. Zhang et al. [14] used convolutional recurrent networks and long-short term memory to separate the near-end speech from the microphone recordings. Fazel et al. [15] proposed deep recurrent neural networks with multitask learning to learn the auxiliary task of estimating the echo in order to improve the main task of estimating the near-end speech.

More recently, the usage of generative adversarial networks in speech enhancement has been studied. Many GAN based speech enhancement algorithms have been proposed. Some are end-to-end solutions that directly mapping a noisy speech to an enhanced signal [16] [17]. Other GANs operate in T-F domain [18] [19], which predict a mask and then resynthesize the target speech in time domain.

In this paper, we propose a GAN-based acoustic echo cancellation algorithm for both linear and nonlinear echo scenarios. In the generator network, log magnitude spectra of microphone signal and reference signal are taken as input, and T-F masks of the spectra are predicted as output. The encoder consists of convolutional layers and decoder comprises deconvolutional layers correspondingly. Between them is a two-layer BLSTM. Convolutional layers are employed to extract the local correlations between microphone signal and reference signal, as well as the mapping relationship between them. BLSTM layers, in the center of G network, can capture long-term temporal information. The discriminator D network has convolutional layers followed by fully connected layers. The input of discriminator is a pair of ground-truth signal and enhanced signal, and output is [0, 1] scaled score instead of True/False.

The remainder of this paper is organized as follows. Section 2 introduces the background knowledge. In Section 3 we present our GAN-based algorithm, followed by experimental setting and results in Section 4. Final conclusion is given in Section 5.

2. Background knowledge

2.1. Acoustic echo cancellation

Acoustic echo is generated by the coupling of a microphone and a speaker, as shown in Figure 1. Far-end signal (or reference signal) $x(t)$ propagate from speaker and through various reflection paths $h(t)$, and is mixed with near-end signal $s(t)$ at the microphone $d(t)$. The acoustic echo is a modified version of $x(t)$ and includes echo path $h(t)$ and speaker distortion.

Conventional AEC algorithms utilize adaptive filter to estimate the echo paths $h(t)$, and subtract the estimated echo $y(t) = \hat{h}(t) * x(t)$ from microphone signal $d(t)$. A separate

double talk detection is required to freeze filter adaption during double talk period. This linear echo canceller is under the assumption of linear relationship between reference signal and acoustic echo. However, nonlinearity is often introduced due to hardware limitation, like speaker saturation. Therefore, a post filter is needed to further suppress the residue echo. The diagram of traditional AEC algorithm is shown as top figure in Figure 2.

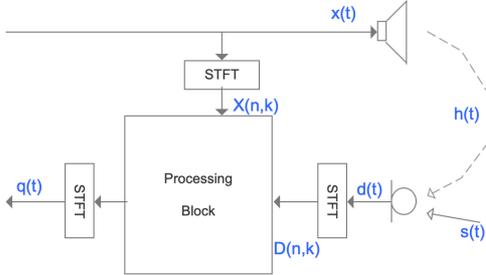


Figure 1 Example of echo generation and acoustic echo cancellation

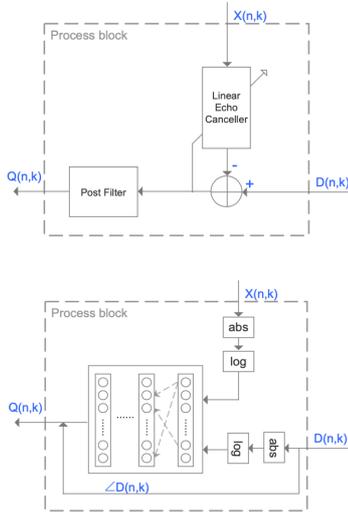


Figure 2 Examples of conventional AEC (top) and neural network based AEC (bottom)

Recently, deep learning based AEC algorithms have shown great potential. With sufficient training data, neural network based solutions yield better performance than traditional ones in both matched and unmatched test cases. Figure 2 bottom shows one example of DNN based algorithm. Model input consists of log magnitude spectra of reference signal and microphone signal. And the model aims to estimate the enhanced magnitude spectrum (similar to LEC and post filter combination). Finally, microphone signal phase is used to resynthesize the enhanced signal in time domain.

General metrics for AEC performance evaluation includes echo return loss enhancement (ERLE) and perceptual evaluation of speech quality (PESQ), which are used in this experiment as well.

ERLE is often used to measure the echo reduction achieve by the system during single talk situation where near-end talker is inactive. ERLE is defined as

$$ERLE(dB) = 10 \log_{10} \frac{E\{d^2(t)\}}{E\{q^2(t)\}} \quad (1)$$

where $E\{\}$ represents the statistical expectation.

PESQ evaluates the perceptual quality of enhanced near-end speech during double talk period. PESQ score is calculated by comparing the enhanced signal to the ground-truth signal, its score ranges from -0.5 to 4.5 and a higher score indicates better quality.

2.2. Generative adversarial network

GAN consists of two networks: a generator network G and a discriminator network D . This forms a minimax game scenario, where G is trying to generate fake data to fool D , and D is learning to discriminate between real and fake data. Importantly, G does not memorize input-output pairs, instead it learns to map the data distribution characteristics to the manifold defined in prior Z . D is typically a binary classifier, and its inputs are either real samples, coming from the dataset that G is imitating, or fake samples, made up by G . As described in [20], the loss function for D and G in conventional GAN can be formulated as

$$\min_G \max_D V(D, G) = \mathbb{E}_{y \sim Y} [\log D(y)] + \mathbb{E}_{z \sim Z} [\log(1 - D(G(z)))] \quad (2)$$

where $\mathbb{E}_{y \sim Y}$ indicates the expectation of y from distribution Y .

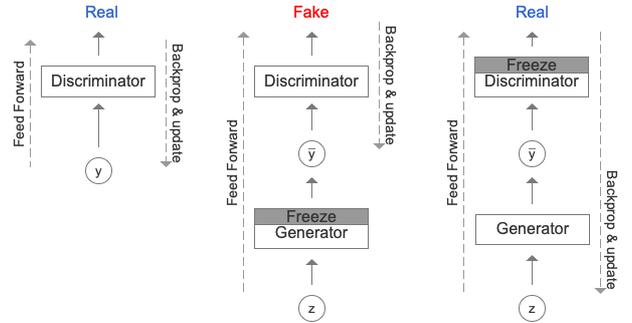


Figure 3 GAN training process

Figure 3 shows the GAN training process. G adapts its parameters such that D classifies G 's output as real. During back-propagation, D gets improved at feeding real features in its input and, in turn, G corrects its parameters to move forwards.

3. Proposed method

The echo cancellation problem is defined so that we have an input echo-corrupted signal $d(t)$ and want to clean it to obtain the enhanced signal $q(t)$. We propose to get this done with GAN model. In the proposed method, the G network performs the enhancement. Its inputs are the log magnitude spectra $D(n, k)$ and $X(n, k)$ of microphone signal $d(t)$ and reference signal $x(t)$ separately, together with the latent representation z , and its output is the T-F masks $Mask(n, k) = G\{D(n, k), X(n, k)\}$ that used to resynthesize the enhanced version $Q(n, k) = Mask(n, k) * D(n, k)$. The G network features an autoencoder-like shape as depicted in Figure 4. In the encoding stage, there are three 2-D convolutional layers follow by a reshape layer. Convolutions enforce the network to focus on temporally-close correlations in the input signal, and were shown to be more stable for GAN training [21]. Correspondingly, the decoding stage is a reversed version of the encoding, which comprises of three deconvolutional layers. Between encoder and decoder, there are two bidirectional

LSTM layers to capture extra temporal information. Batch normalization (BN) [22] is applied after each (de-)convolutional layer except the output layer. Exponential linear units (ELU) [23] are used as activation functions for each layer except output layer that using sigmoid activation function to predict the T-F masks.

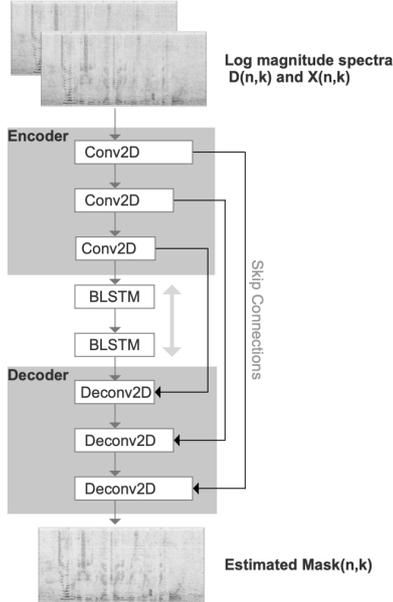


Figure 4 Encoder-decoder architecture of G network

The G network also features skip connections, connecting each encoding layer to its homologous decoding layer, and passing the fine-grained information of the input spectra to the decoder. In addition, they offer a better training behavior, as the gradients can flow deeper through the whole structure [24].

D , on the other hand, is in charge of transmitting information to G of what is real and what is fake, such that G can slightly correct its output towards the realistic distribution, getting rid of the echo components as those are signaled to be fake. D can be expressed as learning some sort of loss for G 's output to look real. D has a similar structure as the encoder in G , where it has three convolutional layers, as well as a flatten layer, then followed by three fully connected layers.

Updating weights towards higher objective metric scores has been proven to work well [19]

$$\begin{aligned} \min_D V(D) &= \mathbb{E}_{(z,y) \sim (z,Y)} [(D(y,y) - Q(y,y))^2] \\ &\quad + \mathbb{E}_{(z,y) \sim (z,Y)} [(D(G(z),y) - Q(G(z),y))^2] \\ \min_G V(G) &= \mathbb{E}_{z,y \sim (z,Y)} [(D(G(z),y) - 1)^2] \end{aligned} \quad (3)$$

where Q stands for the normalized evaluation metric which has its output in range of $[0, 1]$ (1 means the best), and therefore $Q(y,y) = 1$. Furthermore, we found that by adding a L2 norm in $V(G)$ leading to better results:

$$\begin{aligned} \min_G V(G) &= \mathbb{E}_{z,y \sim (z,Y)} [(D(G(z),y) - 1)^2] \\ &\quad + \lambda \|G(z) - Y\|^2 \end{aligned} \quad (4)$$

In the proposed algorithm, log magnitude spectrum is used as input feature. We apply an FFT size of 512 with 25 ms window length and 10 ms step size. PESQ and ERLE are used as the higher objective metrics, and $\lambda = 10$.

Inside the encoder in G , the number of feature maps for convolutional layers are set to: 16, 32, and 64. The kernel size used for the first layer is (1, 3) and for the remaining layers is (2, 3), with strides set to (1, 2). The BLSTM layers consist of 256 neurons, with 128 in each direction and a time step of 100. The decoder part of G follows the reverse parameter setting as in the encoder.

The number of feature maps for convolutional layers in D are set to: 10, 20, 20, and neurons in the fully connected layers are: 30, 10, 1. All models are trained using Adam optimizer [25] for 60 epochs with a learning rate of 0.002 and a batch size of 1. The time step changes with the number of frames per sentence. The input of D is a pair of ground-truth signal and enhanced signal, and output is $[0, 1]$ scaled score instead of True/False. For PESQ loss function, the ground-truth signal is clean near-end speech, and for ERLE loss function, the ground-truth signal is noisy signal (or the mic signal). Both PESQ and ERLE metric loss are based on utterance level.

4. Experimental evaluation

4.1. Experimental settings

TIMIT dataset [26] is used to evaluate the echo cancellation performance. We built a dataset similar to the ones reported in [14] [15]: From 630 speakers of TIMIT, we randomly choose 100 pairs of speakers (40 male-female, 30 male-male, 30 female-female) as the far-end and near-end speakers. Three utterances of the same far-end speaker are randomly chosen and concatenated to create a far-end signal. Each utterance of a near-end speaker is then extended to the same size as that of the far-end signal by zero padding in the rear. Seven utterances of near-end speakers are used to generate 3500 training mixtures where each near-end signal is mixed with five different far-end signals. From the remaining 430 speakers, another 100 pairs of speakers are randomly picked as the far-end and near-end speakers. We followed the same procedure as described above, but this time only three utterances of near-end speakers are used to generate 300 testing mixtures where each near-end signal is mixed with one far-end signal. Therefore, the testing mixtures are from untrained speakers.

The following process is applied to the far-end signal to model the nonlinear acoustic path as in [27]. For the nonlinear model of acoustic path, we first applied the hard clipping to simulate the speaker saturation ($Thr = 80\%$ of the maximum volume of input signal):

$$x_{clip}(t) = \begin{cases} -x_{max}, & \text{if } x(t) < -Thr \\ x(t), & \text{if } |x(t)| \leq Thr \\ x_{max}, & \text{if } x(t) > Thr \end{cases} \quad (5)$$

Then, we applied the following sigmoidal function to simulate the speaker distortion:

$$x_{NL}(t) = 4 \left(\frac{2}{1 + \exp(-a \cdot b(t))} - 1 \right) \quad (6)$$

where $b(t) = 1.5x_{clip}(t) - 0.3x_{clip}(t)^2$, and $a = 4$ if $b(t) > 0$ and $a = 0.5$ otherwise. Finally, the output of sigmoidal function is convolved with a randomly chosen room impulse response (RIR) to render the acoustic echo picked by the microphone.

Image method [28] is used to generate RIRs for training. The length of RIRs is set to 512, the simulation room size is $4m \times 5m \times 3m$, and a microphone is placed at $[2, 2, 1.5]m$. A speaker is placed at five random places with 1.5m distance from

the microphone. The five RIRs' reverberation time (RT_{60}) ranges between 0.2 ~ 0.5 seconds.

Five real environmental recorded RIRs from RWCP database [29] is used to generate acoustic echo in the test. Table I shows the information of the five RIRs.

Table 1 RIRs from RWCP database

RIRs	E1A	E1B	E1C	E2A
RT_{60} (in second)	0.12	0.31	0.38	0.30

Both linear and nonlinear echo scenarios are considered in this test. In training step, microphone signals are generated randomly at signal to echo ratio (SER) {-6, -3, 0, 3, 6} dB, where SER is defined as

$$SER(dB) = 10 \log_{10} \frac{E\{signal_{near}^2\}}{E\{signal_{far}^2\}} \quad (7)$$

In testing stage, microphone signals are generated at SER levels {0, 3.5, 7} dB, slightly different from the training SERs, in order to evaluate the unmatched training-test cases.

4.2. Experimental results

In this experiment, two state-of-the-art neural network algorithms, CRNN [14], and Multitask GRU [15] are deployed as the benchmark. It is shown that NN based methods outperforms conventional 'AES+RES' methods in [15], and thus we skip the 'AES+RES' comparison here. Instead of original implementation in [14], we directly used the G in the proposed method as CRNN structure, and the parameters are given in Section 3. Multitask GRU was implemented following the instructions in [15].

Table 2 PESQ and ERLE scores in linear acoustic path scenarios

Metrics	Methods	Testing SER (dB)		
		0	3.5	7
PESQ	Input	1.80	2.02	2.28
	CRNN	2.67	2.80	3.18
	Multitask GRU	2.84	3.05	3.30
	CRGAN-EC-P	2.90	3.10	3.46
	CRGAN-EC-E	2.54	2.72	3.06
ERLE (dB)	Input	0	0	0
	CRNN	59.60	60.66	61.10
	Multitask GRU	58.20	59.14	59.60
	CRGAN-EC-P	59.67	60.50	61.21
	CRGAN-EC-E	64.12	64.10	67.28

We first evaluate our proposed method in linear acoustic path scenarios. Table 2 shows the average PESQ and ERLE scores of the unprocessed, CRNN, Multitask GRU, and the proposed methods. 'CRGAN-EC-P' denotes the proposed GAN using PESQ as metric loss, and 'CRGAN-EC-E' denotes GAN with ERLE loss correspondingly. The results show that CRGAN-EC-P yields the best PESQ scores among all the methods, and comparable ERLE scores to CRNN and Multitask GRU. The superior performance of CRGAN-EC-P could be explained as following. Convolutional layers contribute in providing local correlations and mapping between the microphone signal and reference signal, and BLSTM layers

contribute in capturing long-term temporal information. Furthermore, traditional mean square error (MSE) loss function measures the spectrum distance between enhancement signal and ground-truth signal with uniform weights, while PESQ is an accumulated score along subbands with different weights according to psychoacoustics. Therefore, models minimizing MSE score do not guarantee to yield good PESQ score.

CRGAN-EC-E obtains worse PESQ scores compared to CRNN and Multitask GRU, but it produces highest ERLE scores. Since noise is not considered and SER is not low in this experiment, high ERLE score does not make much sense. However, different metric losses provide additional freedom for the system to achieve specific goals. And metric loss of combined PESQ and ERLE could further achieve appropriate compromise between echo cancellation and target speech distortion. Due to limited space, only separate PESQ and ERLE results are shown here.

We further study the impact of nonlinear acoustic path on our proposed method. In this test, x_{NL} was convolved with RIRs to generate the acoustic echo, and thus it contained both power amplifier clipping, and loudspeaker distortion. We again compare results of our method against CRNN and Multitask GRU. Table 3 shows the averaged PESQ and ERLE scores. Similarly, the Proposed CRGAN-EC-P achieves the best PESQ scores among all the methods and the ERLE scores are close to the best ones. Proposed CRGAN-EC-E yields the best ERLE scores.

Table 3 PESQ and ERLE scores in nonlinear acoustic path scenarios

Metrics	Methods	Testing SER (dB)		
		0	3.5	7
PESQ	Unprocessed	1.68	1.91	2.18
	CRNN	2.38	2.69	2.94
	Multitask GRU	2.54	2.80	3.11
	CRGAN-EC-P	2.75	2.99	3.32
	CRGAN-EC-E	2.30	2.54	2.88
ERLE (dB)	Unprocessed	0	0	0
	CRNN	57.14	56.99	58.30
	Multitask GRU	55.30	56.23	57.02
	CRGAN-EC-P	57.01	57.89	58.26
	CRGAN-EC-E	62.71	64.30	66.93

5. Conclusion

In this study, we proposed a novel acoustic echo cancellation algorithm using convolutional recurrent GAN, that works well in both linear and nonlinear acoustic path scenarios. Convolutional layers extract the local correlations between microphone signal and reference signal, and the mapping between them. BLSTM layers capture the long-term temporal information. The proposed architecture is trained in frequency domain and predicts the time-frequency (TF) mask for the target speech. We deploy various metric loss functions and demonstrate the model robustness to the trade-offs between echo suppression and target speech distortion. As future works, we intend to extend our investigation of echo cancellation under more severe situations, especially delay-free scenarios.

6. References

- [1] E. Hansler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, Adaptive and learning systems for signal processing, communications, and control, Hoboken, NJ: Wiley-Interscience, 2004.
- [2] S. Gustafsson, R. Martin and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, vol. 64, no. 1, pp. 21-32, 1998.
- [3] V. Turbin, A. Gilloire and P. Scalart, "Comparison of three postfiltering algorithms for residual acoustic echo reduction," in *Proc. IEEE International Conf. on Acoustic, Speech, and Signal Processing*, 1997.
- [4] A. Stenger, L. Trautmann and R. Rabenstein, "Nonlinear acoustic echo cancellation with 2nd order adaptive Volterra filters," in *Proc. IEEE International Conf. on Acoustics, Speech, and Signal Processing*, 1999.
- [5] G. Hinton, L. Deng, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, 2012.
- [6] X. Lu, Y. Tsao, S. Matsuda and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Conf. of International Speech Communication Association*, 2013.
- [7] Y. Xu, J. Du, L. Dai and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65-68, 2014.
- [8] Y. Wang, A. Narayanan and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849-1858, 2014.
- [9] F. Weninger, J. Hershey, J. Roux and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Global Conf. on Signal and Information Processing*, 2014.
- [10] C. Lee, J. Shin and N. Kim, "DNN-based residual echo suppression," in *Proc. Conf. of International Speech Communication Association*, 2015.
- [11] M. Muller, J. Jansky, M. Bohac and Z. Koldovsky, "Linear acoustic echo cancellation using deep neural networks and convex reconstruction of incomplete transfer function," in *IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics*, 2017.
- [12] H. Zhang and D. Wang, "Deep learning for acoustic echo cancellation in noisy and double-talk scenarios," in *Proc. Conf. of International Speech Communication Association*, 2018.
- [13] G. Carbajal, R. Serizel, E. Vincent and E. Humbert, "Multiple-input neural network-based residual echo suppression," in *IEEE International Conf. on Acoustic, Speech and Signal Processing*, 2018.
- [14] H. Zhang, K. Tan and D. Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions," in *Proc. Conf. International Speech Communication Association*, 2019.
- [15] A. Fazel, M. El-Khamy and J. Lee, "Deep multitask acoustic echo cancellation," in *Proc. Conf. International Speech Communication Association*, 2019.
- [16] S. Pascual, A. Bonafonte and J. Serra, "Segan: Speech enhancement generative adversarial network," in *arXiv preprint*, 2017.
- [17] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *Proc. International Conf. on Acoustic, Speech and Signal Processing*, 2019.
- [18] M. Soni, N. Shah and H. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *Proc. International Conf. on Acoustic, Speech and Signal Processing*, 2018.
- [19] S. Fu, C. Liao, Y. Tsao and S. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *arXiv preprint*, 2019.
- [20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," in *Conf. on Neural Information Processing Systems*, 2014.
- [21] A. Radford, L. Metz and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *arXiv preprint*, 2015.
- [22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *arXiv preprint*, 2015.
- [23] D. Clevert, T. Unterthiner and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," in *arXiv preprint*, 2015.
- [24] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. on Computer Vision and Pattern*, 2016.
- [25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv preprint*, 2014.
- [26] F. Lamel, R. Kassel and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Speech Input/Output Assessment and Speech Databases*, 1989.
- [27] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2065-2079, 2012.
- [28] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of Acoustic Society of America*, vol. 65, no. 4, pp. 943-950, 1979.
- [29] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. International Conf. on Language Resources and Evaluation*, 2000.