GSC Based Speech Enhancement with Generative Adversarial Network

Yao Zhou, Changchun Bao, Rui Cheng
Speech and Audio Signal Processing Laboratory, Faculty of Information Technology,
Beijing University of Technology, Beijing, 100124, China
E-mail: zhouyao@emails.bjut.edu.cn, baochch@bjut.edu.cn, chengrui@emails.bjut.edu.cn

***Abstract*— At present, the technology of using microphone arrays for speech enhancement has been widely concerned, and the enhancement effect is excellent. The widely used Generalized Sidelobe Canceller (GSC) method can achieve good noise reduction for noisy speech in the additive noise acoustic environment, and achieve better intelligibility improvement. But there are also areas for improvement. In the lower branch of GSC, signal leakage caused by the estimation of the incident angle or the slight change of the position of the microphone array may cause the self-cancellation of target speech signal, thereby the severe speech distortion is caused. In this paper, the Generative Adversarial Network (GAN), which has broad application prospects in deep learning technology, replaces the lower branch of the traditional GSC structure, thus the self-cancellation of speech signals is avoided and improving the anti-error ability of the enhancement system is improved effectively.**

I. INTRODUCTION

There is a problem in almost speech enhancement methods, while dealing with the single-channel speech enhancement, that is, the distortion of the speech signal varies with the input signal-to-noise ratio (SNR) to some extent. Although the quality of speech could be improved by the methods, the speech intelligibility is seriously affected. If we adopt a microphone array technology that can take advantage of the spatial characteristics of the sound source, the speech distortion could be theoretically avoided as much as possible, while improving the speech quality and SNR of the enhanced speech, thereby the subsequent processing would be better served [1]. Presently, the microphone array technologies have been well applicated in speech recognition and speech enhancement. At the same time, the sound source localization technology based on microphone array is also widely used in practical applications such as target monitoring and tracking, teleconference, and hearing aids.

Nowadays, the field of machine learning is developing rapidly and has gone deep into our daily life. Speech recognition, image recognition and machine translation are all changing our lifestyle, and these are all closely related to deep learning. Its purpose is to construct a deep neural network which can carry out adaptive analysis learning. Its essence is to layer feature representation of input data, further abstract low-level features into high-level features, and then learn the mapping relationship between high-level features and objectives, so that in the testing stage, it can be separated from the data set and achieve certain prediction and estimation capabilities [2]. And generation adversarial network (GAN), proposed in [3], has promoted deep learning to a new height, and its various applications is emerging endlessly. Currently, it is the popular neural network in the deep learning area.

Combining traditional microphone array signal processing technology with deep neural network, the advantage of deep neural network in data fitting ability is brought into full play, and the speech enhancement technology based on microphone array is further improved. In the early stage of the development of neural network [4], the author proposed a speech enhancement method based on neural network. In [5], the author proposed a method of realizing speech enhancement with generative adversarial networks (GAN), named SEGAN. It inspired this paper's work.

In this paper, the generalized sidelobe canceller (GSC) beamformer [6] is mentioned as a popular microphone array speech enhancement method. However, this method has a defect which limits its application. It is sensitive to time difference of arrival (TDOA), thus it has a high dependence on the estimation of the array structure and the incident angle.

The main purpose of this paper is to improve the TDOA-sensitive defect of the GSC by utilizing the GAN, and the improved system is called GAN-GSC in this paper. The GAN-GSC would be used to achieve the enhancement of noisy speech signals in the additive noise acoustic environment, and improve the subjective auditory quality. Then, segmented signal to noise ratio (SSNR) [7], short-time objective intelligibility (STOI) [8] and perception evaluation of speech quality (PESQ) [9], are used as evaluation metrics to test whether the new speech enhancement system performs well.

The rest of this paper is organized as follows. In Section 2, the traditional GSC method is reviewed. In Section 3, the details of the proposed GAN-GSC method are discussed. Experiments and results are provided in Section 4 & 5, and the conclusions are given in Section 6.

II. REVIEW OF GENERALIZED SIDELOBE CANCELLER

The GSC was proposed by Griffiths [6], which transforms the linearly constrained minimum-variance (LCMV) proposed by Frost [10] into an unconstrained situation. As can be shown in Fig. 1, $[y_1(n), y_2(n), \ldots, y_M(n)]$ is assumed as the speech inputs of the M microphones. And the structure of the GSC is mainly composed of two subsystems, the first subsystem is the fixed beamformer in the upper branch consisting of a delay alignment structure and a set of fixed weights $\boldsymbol{W}_c = [w_{c1}, w_{c2}, \ldots, w_{cM}]$. By adjusting $\boldsymbol{W}_c$, the gain and beam width of

the fixed beamformer can be controlled. The second subsystem is the blocking matrix $\boldsymbol{B}$ in the lower branch, which used to estimate the noise component. The output of the two subsystems are used for adaptive cancellation by adaptive canceller with a coefficient matrix $\boldsymbol{W_a}$ to get the enhanced speech. In [6], traditional least mean square (LMS) algorithm [11] is used to obtain the adaptive weights $\boldsymbol{W_a}$. It is clear that $y_a(n)$ is the output of the fixed beamformer, where $y_z(n)$ denotes the noise component in $y_a(n)$, and $y_o(n)$ is the enhanced speech.

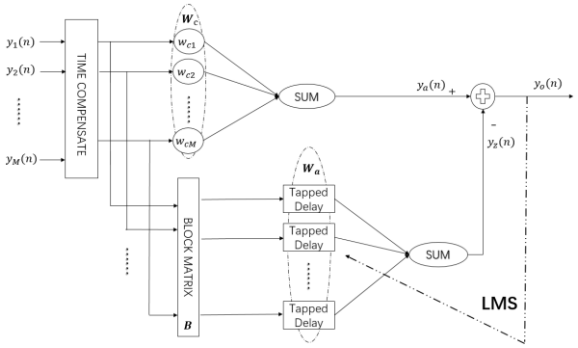


Fig. 1 The structure of the GSC

In this paper, the fixed beamformer with delay-and-sum (DS) [12] is used in the upper branch. And the blocking matrix in the lower branch is designed as follows

$$\boldsymbol{B}^H = \begin{bmatrix} 1 & -1 & 0 & \cdots & & 0 \\ 0 & 1 & -1 & \cdots & & 0 \\ & \vdots & & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & & -1 \end{bmatrix} \tag{1}$$

The adaptive cancellation process of the two outputs of the system can be expressed as

$$\boldsymbol{W_a} = \min_{\boldsymbol{W_a}} \{[\boldsymbol{y_a} - \boldsymbol{B}\boldsymbol{W_a}]^H \boldsymbol{R}_{y_o y_o} [\boldsymbol{y_a} - \boldsymbol{B}\boldsymbol{W_a}]\} . \tag{2}$$

where $\boldsymbol{y_a}$ represents the output of the fixed beamformer, $\boldsymbol{R}_{y_o y_o}$ represents the autocorrelation matrix of final output. In order to avoid computational complexity of derivative operation, the normalized least mean square (NLMS) [13] error is used for approximation, i.e.

$$\boldsymbol{W_a}(n+1) = \boldsymbol{W_a}(n) + \mu(n) e(n) \boldsymbol{B}^H \boldsymbol{y}(n) \tag{3}$$

$$\mu(n) = \frac{\beta}{P_x + \alpha}, P_x = (\boldsymbol{B}^H \boldsymbol{y}(n))^T \boldsymbol{B}^H \boldsymbol{y}(n) . \tag{4}$$

where $\boldsymbol{y}(n)$ denotes multichannel inputs, $\mu(n)$ means the variable step size of the NLMS, and $e(n)$ corresponds to the final outputs $y_o(n)$ shown in Fig. 1. It can be deduced that the $\boldsymbol{B}^H\boldsymbol{y}(n)$ means the input of adaptive process, and P_x is the energy of input data. β is a positive real scale factor, and α is a small value to avoid infinite $\mu(n)$ when P_x is zero.

From the above deduction, we can find that if the compensation error of time delay exists, the multi-channel signals are not aligned. When the signals pass through the blocking matrix, there will be a correlation component with the pure speech in the residual signal. This will cause the cancellation of the pure speech in the fixed beamforming output of the upper branch in the multi-channel canceller. The phenomenon of signal self-cancellation makes the signal seriously distorted and the enhancement quality degraded.

III. GSC BEAMFORMING BASED ON THE GAN

A. Overview of the GAN-GSC

In this paper, in order to overcome the shortage of the GSC, the generator of the GAN is used to replace lower branch portion of the GSC, as shown in Fig. 2. The upper branch is the same as the GSC, i.e., a DS beamformer. The generator of the GAN is used to estimate $\hat{y}_z(n)$, that is the noise component in $y_a(n)$. By a subtraction operation, the final enhanced output of the system $y_o(n)$ can be obtained. The proposed structure of the GSC beamforming based on the GAN is named as GAN-GSC.

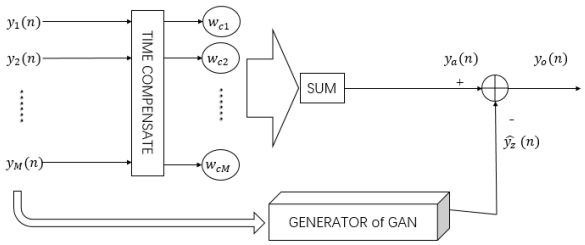


Fig. 2. The structure of the GAN-GSC speech enhancement system

B. The Proposed GAN

The traditional GAN is a type of generation model with two modules in its network, one of them is *G* (Generator) network. The other one is *D* (Discrimination) network. This model greatly absorbs the idea of game theory, and the two networks play each other for the desired effect. The *D* network intents to be able to correctly discriminate whether the input data is from a real sample or a forged sample generated by a *G* network through continuous training. The purpose of the *G* network is to generate data that is very close to the real sample through continuous training, so that the *D* network cannot distinguish the authenticity [3]. The mathematical description of this process can be expressed as

$$\min_G \max_D V(G,D) = E_{x\sim P_{data}(x)}[\log(D(x))] + E_{z\sim P(z)}[1-\log D(G(z))] . \tag{5}$$

The cross-loss entropy is used as loss function of the traditional GAN. In order to make the network easier to achieve balance and get more accurate estimation results, the loss-sensitive generative adversarial network (LSGAN) proposed in [14] is considered in this paper. The loss function is changed into least squares, so the loss functions of *D* and *G* are shown in (6) and (7), respectively.

$$\min_D V_{ls}(D) = 0.5E_{x\sim P_{data}(x)}[(D(x)-1)^2] + 0.5E_{z\sim P(z)}[(D(G(z)))^2] . \tag{6}$$

$$\min_G V_{ls}(G) = 0.5E_{z\sim P(z)}[(D(G(z))-1)^2] . \tag{7}$$

The input matrix is related to three values, namely *M*, *K*, and *2*. The digit *2* denotes two items, $A(n,k)$ and $\Psi(n,k)$, which

are the amplitude and phase of short-time Fourier transform (STFT) of the noisy multichannel speech $\boldsymbol{y}(n)$. Here, n is the frame index, k is the frequency bin index. The M is the number of microphones, and the K is the total number of frequency bins. So, considering the idea in [15], the convolutional neural networks (CNN) structure is used in the generator (G) to capture the high-level information from input features, shown in the Fig. 3. The size of the convolutional kernel is 2×1, and the strides is 1. In order to better extract spatial information of input multi-channel features, the input matrix, convolution kernel and strides are all small. LeakReLU [16] is used as the active function for all layers. Dropout regularization [17] with a probability of 0.8 is utilized in full connected layers.

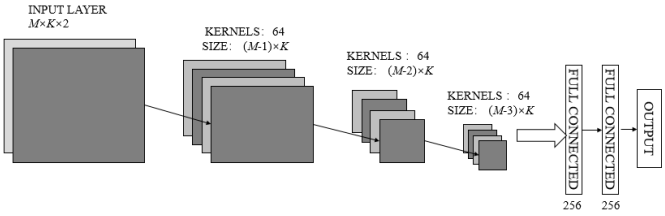


Fig. 3. Generator structure

In order to enhance speech more effectively, we use mask as the target of the G network. So, both the ideal mask and the estimated mask are applied as inputs of discriminator (D) network. Depicted in Fig. 4, it consists of four full connected layers, with dropout regularization and LeakReLU active function. At the final, sigmoid [18] function is employed in the output layer.

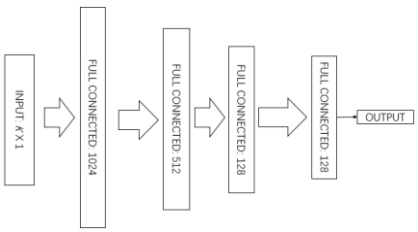


Fig. 4. Discriminator structure

C. Speech Enhancement Based on the GAN-GSC

In this paper, the proposed GAN is used for GSC-based speech enhancement. For obtaining the estimated value of the required noise component, a method based on Time-Frequency (T-F) mask is used. CNN-based generator G in the GAN is used to estimate the ideal ratio mask (IRM) of noise components. Its mathematic definition is proposed in [19], i.e.

$$M_{IRM}(n,k)=\frac{\left|Y_z(n,k)\right|}{\left|S(n,k)+Y_z(n,k)\right|}. \tag{8}$$

where $S(n,k)$ is the STFT of the desired speech, and $Y_z(n,k)$ is the STFT of the noise signal in the output of fixed beamformer.

So, the STFT of noise components could be obtained by multiplying $M_{IRM}(n,k)$ with the STFT value $Y_a(n,k)$ of the output of fixed beamformer. Then, doing an inverse STFT, the estimated target noise component $\hat{y}_z(n)$ is obtained as follows.

$$\hat{Y}_z(n,k)=\hat{M}_{IRM}(n,k)Y_a(n,k). \tag{9}$$

By subtracting the estimated noise signal $\hat{y}_z(n)$ from the fixed beamformer outputs $y_a(n)$, the enhanced speech signal $\hat{s}(n)$ is given as follows.

$$\hat{s}(n)=y_a(n)-\hat{y}_z(n). \tag{10}$$

IV. The Setting of the Data

A. The Signal Model

As can be shown in Fig.5, a uniform linear array of 16 unidirectional microphones is applied in this paper for speech enhancement in the condition of far-field single-target speech sources with additive noise and no reverberation acoustic model.

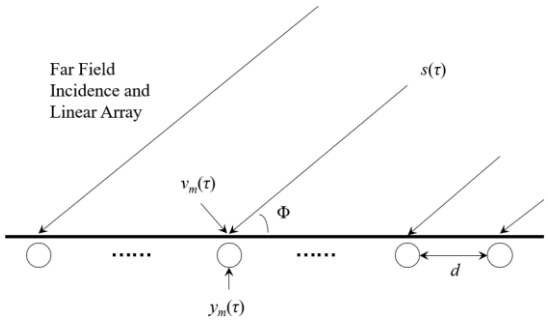


Fig. 5. Uniform linear microphone array and Far-field incidence structure.

The matheamtical modeling of this acoustic structure is described in (11). $s(n)$ denotes the target speech, $v_m(n)$ denotes the additive noise of the m^{th} microphone, and $y_m(n)$ denotes the noisy input signal of the m^{th} microphone, Φ is the incident angel, $F_m(\Phi)$ denotes the TDOA, t denotes the fixed propagation delay, d denotes the distance between the adjacent microphones.

$$y_m(n)=S[n-t-F_m(\Phi)]+v_m(n). \tag{11}$$

The most prominent feature of the GAN is that it can achieve unsupervised learning and get target output by using unlabeled noise input. In the experiment of this paper, the input is tagged, which is more similar to the conditional GAN. Considering that the input and output of the GAN in this paper are noise components, the conditional information of input term in the classical GAN with the condition is omitted. Here, the discriminant network part can be understood as a loss function requiring the training.

B. The Setting of Acoustic Environment

The direction of incidence of the speech source is assumed to arrange from 70° to 110°, indicating that the speaker is facing the microphone array, white noise is set at 60°, pink noise is presumed at 150°, and Babble noise is used to denote the voice of discussion existing at the directions from 0° to 180°. Mix all the noises together and apply it to the microphone in

accordance with the specified SNR. Three kinds of input SNR levels (0dB, 5dB, and 10dB) are considered.

C. The Setting of the Training and Test Data

As a widely used speech database TIMIT corpus [20] is applied to experiments. And at the same time, the noise to be used will be obtained from the NOISE92 library [21]. The incident angles of 40 speech segments randomly chosen from TIMIT, are chosen from 70° to 110° in step of 5°. Simultaneously, one of 6 angles evenly from 0° to 180° are selected as the position of Babble noise.

While generating test data, incident angles and speech segments all are chosen randomly for better detection the practicality of the enhancement system.

V. Experiments and Results

After training the neural network to a good performance, a series of tests are conducted to observe the enhancement effect of the GAN-GSC.

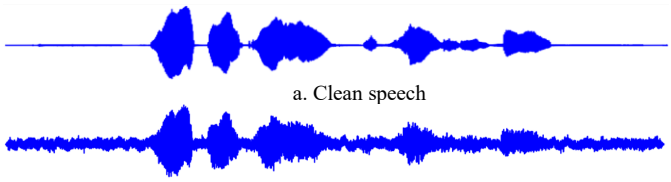
A. The Comparison of the GAN-GSC and GSC

This experiment will compare the performance of the traditional GSC (named as GSC) and GAN-GSC. Three segments of noisy multi-channel speech are randomly selected from the test set, and then, enhanced by the GAN-GSC and GSC, respectively. Table 1 shows the average results of the SSNR [7], PESQ [9] and STOI [8] performance, the last column in the Table.1 shows the improvement of the proposed GAN-GSC versus the GSC.

Table. 1. The average results of the quality and intelligibility test

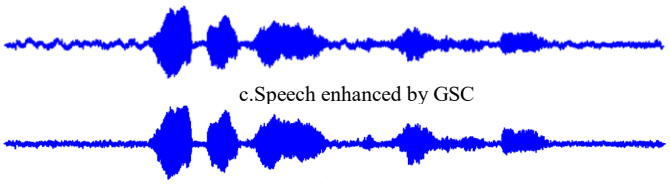
	SNR	Noisy	GSC	GAN-GSC	Impro.
SSNR	0	-19.308	4.744	5.211	0.467
	5	-14.298	2.968	4.338	1.370
	10	-9.337	1.005	4.433	3.428
PESQ	0	1.532	2.223	2.258	0.035
	5	1.877	2.456	2.639	0.183
	10	2.248	2.643	2.946	0.303
STOI	0	0.696	0.815	0.821	0.006
	5	0.803	0.874	0.881	0.007
	10	0.896	0.924	0.934	0.010

For better observation of the enhancement result, speech waveform and spectrograms (SNR=5dB) are shown in Fig. 6 and Fig.7, respectively. Here, all waveforms in Fig. 6 are normalized.



a. Clean speech

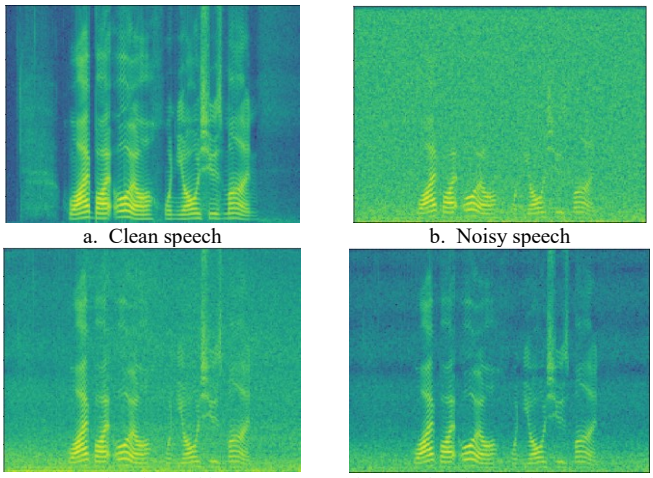
b. Noisy speech



c.Speech enhanced by GSC

d. Speech enhanced by GAN-GSC

Fig. 6 Speech waveform comparison



a. Clean speech b. Noisy speech

c. Speech enhanced by GSC d. Speech enhanced by GAN-GSC

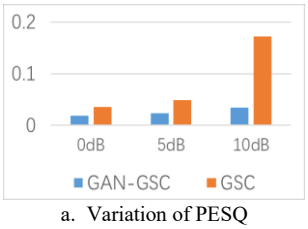
Fig. 7 Spectrogram comparison

From the Table. 1, Fig. 6 and Fig. 7, obviously, the quality and intelligibility of the speech enhanced by the proposed method are higher than those enhanced by the traditional GSC method. It can also be found that as the SNR increases, the improvement of the quality and intelligibility are more obvious. This is because the higher the SNR, the more obvious the characteristics of the speech signal, and the easier it is to learn for neural networks.

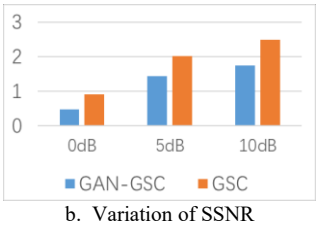
B. Error Tolerance Test

For testing effectiveness of the proposed method for the shortage of the GSC, the incident angle errors are artificially set. The variation of the quality parameters of the speech enhanced by the two methods in the presence or absence of error is compared.

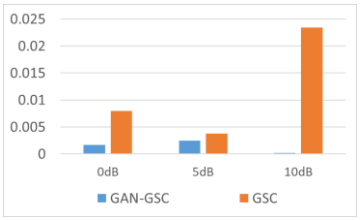
With randomly choosing speech segment and setting a 5° incident angel error, the variations of the SSNR, PESQ and STOI can be clearly seen in Fig. 8.



a. Variation of PESQ



b. Variation of SSNR



c. Variation of STOI

Fig. 8 Error tolerance test results

In Fig. 8, the orange color column denotes the GSC, and the blue color column denotes the GAN-GSC. It is obvious to see that orange one is much higher than blue one. Compared to the GSC, the quality of speech enhanced by the GAN-GSC generates slighter self-cancellation with the incident angel error. This means that the GAN-GSC has better error tolerance than the GSC.

In order to more intuitively observe the impact of incident angle error on the speech enhanced by the two methods, with a 5° error, the leak waveform of speech shows how serious the speech self-cancellation is, that is, the enhanced speech signal subtracted from the clean speech signal, is given below.

a. Leak waveform of speech enhanced by the GSC

b. Leak waveform of speech enhanced by the GAN-GSC

Fig. 9 Leak waveform comparison

It is clear that in Fig. 9b, the speech components are much less than Fig. 9a. Thus, the enhancement performance of the GAN-GSC is much more stable. And the little residual of speech components in Fig. 9b might be caused by the attenuation of the fixed beamformer. Please note that the waveforms shown in Fig. 9 have been normalized.

At the same time, by contrasting the noise components in the enhanced speech in Fig. 9a and Fig. 9b, the noise reduction capacity of the GAN-GSC is absolutely better, whether there are incident angel errors or not.

VI. CONCLUSIONS

Microphone array speech enhancement technology was mainly studied in this paper. In the simulated additive noise acoustic environment, an adaptive beamformer, GSC is build. And then, aiming at the defects of the GSC, combining with the deep neural network technology, the traditional structure was improved by the GAN, and successfully achieved the enhancement of the noisy speech under different SNR levels, as well as the improvement of the subjective intelligibility. In general, it not only improved the speech enhancement performance of the traditional structure, but also better avoided the phenomenon of target speech self-cancellation in the GSC.

But the corresponding problems still exist, first of all, the fixed beamforming of the upper branch in the GSC structure largely limits its performance. So, in the future, the structure of the upper branch in the GSC can be further improved or replaced it with other adaptable beamforming methods to obtain a better performance. In addition, dereverberation in speech enhancement is a very difficult problem. This paper only studied the additive noise environment and improved the ability of dealing with reverberation is insufficient. In the future research, the reverberation will also be taken into account.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (Grant No. 61831019, No. 61471014 and No. 61231015).

REFERENCES

[1] Benesty, Jacob , J. Chen , and Y. Huang . "Microphone Array Signal Processing." (2008).
[2] Hagan, Martin T , H. B. Demuth , and M. H. Beale. "Neural network design." China Machine Press, 2002.
[3] Goodfellow, Ian J , et al. "Generative Adversarial Nets." International Conference on Neural Information Processing Systems MIT Press, 2014.
[4] Parveen, S. , and P. Green . "Speech enhancement with missing data techniques using recurrent neural networks." IEEE International Conference on Acoustics IEEE, 2004.
[5] Pascual S , Bonafonte A , Serrà, Joan. "SEGAN: Speech Enhancement Generative Adversarial Network." Interspeech, 2017, 3642-3646.
[6] Griffiths L J. "An Alternative Approach to Linear Constrained Adaptive Beamforming." IEEE transactions on Antennas and Propagation, 1982, 30(1):27-34.
[7] John H L, Bryan H, Pellom L, "An Effective Quality Evaluation Protocol For Speech Enhancement Algorithms."International Conference on Speech & Language Processing, 1988, pp. 2819-2822.
[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] "Perceptual evaluation of speech quality (PESQ): An objective method for endto-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU Rec. P.862, 2001.
[10] Frost, O. L , III. "An algorithm for linearly constrained adaptive array processing." Proceedings of the IEEE, 1972, 60(8):926-935.
[11] B. Widrow and M. E. Hoff, "Adaptive switching circuits," Neurocom- puting: foundations of research, pp. 123-134, 1988.
[12] M. Brandstein and D. B. Ward, eds., Microphone Arrays: Signal Processing Techniques and Applications. Berlin, Germany: Springer-Verlag, 2001.
[13] Rupp, Markus . "The Behavior of LMS and NLMS Algorithms in the Presence of Spherically Invariant Processes." IEEE Transactions on Signal Processing 41.3(1993):1149-1160.
[14] Qi, G.J., "Loss-Sensitive Generative Adversarial Networks on Lipschitz Densities." 2017, arXiv preprint arXiv:1701.06264.
[15] Chakrabarty Soumitro , Wang DeLiang , Habets Emanuël. "Time-Frequency Masking Based Online Speech Enhancement with Multi-Channel Data Using Convolutional Neural

Networks." International Workshop on Acoustic Echo and Noise Control. Tokyo, Japan, 2015: 476-480.
[16] Xu, Bing , et al. "Empirical Evaluation of Rectified Activations in Convolutional Network." Computer Science, 2015.
[17] Wahlbeck K, Tuunainen A, Ahokas A, et al. "Dropout rates in randomised antipsychotic drug trials." Psychopharmacology, 2001, 155(3):230-233.
[18] Ito, Yoshifusa . "Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory." Neural Networks 4.3(1991):385-394.
[19] Wang Y , Narayanan A , Wang D L . "On Training Targets for Supervised Speech Separation." IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12):1849-1858.
[20] Garofolo J S, Lamel L F, Fisher W M, et al. "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1." Nasa Sti/recon Technical Report N, 1993, 93.
[21] A. Varga, and H. J. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” Speech communication, vol. 12, no. 3, pp. 247-251, 1993.