# REMIX-CYCLE-CONSISTENT LEARNING ON ADVERSARIALLY LEARNED SEPARATOR FOR ACCURATE AND STABLE UNSUPERVISED SPEECH SEPARATION

*Kohei Saijo, Tetsuji Ogawa*

Department of Communications and Computer Engineering, Waseda University, Tokyo, Japan

## ABSTRACT

A new learning algorithm for speech separation networks is designed to explicitly reduce residual noise and artifacts in the separated signal in an unsupervised manner. Generative adversarial networks are known to be effective in constructing separation networks when the ground truth for the observed signal is inaccessible. Still, weak objectives aimed at distribution-to-distribution mapping make the learning unstable and limit their performance. This study introduces the remix-cycle-consistency loss as a more appropriate objective function and uses it to fine-tune adversarially learned source separation models. The remix-cycle-consistency loss is defined as the difference between the mixed speech observed at microphones and the pseudo-mixed speech obtained by alternating the process of separating the mixed sound and remixing its outputs with another combination. The minimization of this loss leads to an explicit reduction in the distortions in the output of the separation network. Experimental comparisons with multichannel speech separation demonstrated that the proposed method achieved high separation accuracy and learning stability comparable to supervised learning.

***Index Terms*—** Deep neural networks, adversarial learning, remix-cycle-consistent learning, unsupervised speech separation

## 1. INTRODUCTION

Since speech recordings are frequently contaminated by interference and background noise, source separation technology has become essential in voice applications. Deep neural networks (DNNs) have been shown to give remarkably high separation performance [1–3], especially in supervised settings where models are trained using large number of pairs of observed signals and corresponding ground truths (i.e., distortion-free signals). It, however, is not feasible to collect such distortion-free data in real environment, and learning source separation models under unsupervised conditions where paired data are not available is thus highly desired.

Adversarial learning [4] of source separation models aims to bring the distribution of separated signals closer to the distribution of distortion-free signals. It was originally introduced in supervised settings [5–10] but is more potent in unsupervised situations where the ground truth is not given [11–16]. However, its weak objectives aimed at distribution-to-distribution mapping destabilize the learning (i.e., performance is highly dependent on initial parameters) and limit the separation performance. Several attempts have been made to overcome these shortcomings by exploiting the remix-cycle-consistency loss [11] and the boundary equilibrium generative adversarial network [12].

In contrast, this study attempts to solve the above problem by imposing a more direct requirement that reduces residual noise and artifacts in the separated signal on unsupervised learning of source separation models. For this purpose, a two-stage learning algorithm is proposed: the neural separator (specifically, a minimum variance distortionless response (MVDR) beamformer [17] with DNN-based time-frequency (TF) masking [18–20]) is adversarially trained and then fine-tuned using remix-cycle-consistency loss. This loss is defined as the difference between the mixed sound observed at microphones and the pseudo-mixed sound obtained by alternating the process of separating the mixed sound and remixing its outputs with another combination. Since the repetitive process of unmixing and remixing accumulates distortions, minimizing the remix-cycle-consistency loss leads to the removal of residual noise and artifacts in the separator's output. In this case, using the observed mixture as a teacher means imposing constraints on learning at the input speech level rather than at the distribution level. Therefore, tuning the separator to minimize this loss is expected to improve the separation accuracy and reduce the instability of adversarial learning.

There are two key contributions made in this work: 1) we provide an algorithm that simultaneously addresses the performance limitation and learning instability of neural separators adversarially learned in an unsupervised manner, and 2) we provide insights into the effects of using remix-cycle-consistent learning in the latter stage of two-stage learning: this would be appropriate to be used for fine-tuning of the neural separators.

The remainder of this paper is organized as follows. Section 2 provides the elemental technologies used. Section 3 describes the details of the proposed two-stage learning algorithm for stable and accurate speech separation. Section 4 demonstrates the effectiveness of the proposed method through experiments. Section 5 concludes this paper.

## 2. ELEMENTAL TECHNOLOGY

Let us denote mixtures of $N$ sources captured by $M$ microphones in the short-time Fourier transform (STFT) domain as $\boldsymbol{x}_{f,t} \in \mathbb{C}^M$, where $f$ and $t$ denote a frequency bin and a time index, respectively, and are omitted hereafter for simplicity.

### 2.1. Adversarial Learning

GANs [4] consist of two modules, a generator and a discriminator. The discriminator tries to distinguish real data from the generator's output (i.e., fake data), while the generator tries to yield data that can fool the discriminator, which leads to the generator that outputs data similar to real data. Optimization of GANs is formulated as:

$$\min_G \max_D \mathcal{L}_{\mathrm{GAN}} = \mathbb{E}_{\boldsymbol{y}\sim p_{\mathrm{data}}(\boldsymbol{y})}[\log D(\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}}(\boldsymbol{x})}[\log\left(1 - D(G(\boldsymbol{x}))\right)], \quad (1)$$

where $G$ and $D$ denote the generator and the discriminator, and $\boldsymbol{y}$ and $\boldsymbol{x}$ represent the real and the fake, respectively. This equation implies that the generator is trained to make the distribution of its output closer to the real data distribution.

When applying GANs to speech separation, the generator is the separator to be sought, and the real data are distortion-free (clean) speech signals. GANs were originally introduced for speech separation in a supervised setting where the ground truth for each observed mixture could be given [5]. This framework, however, is more potent in unsupervised settings [12, 13] where any clean signal that is not paired with an observed mixture can be used as real data, because its objective is based on distribution-to-distribution mapping.

### 2.2. Adversarial Unmix-and-Remix

An attempt was made to exploit cycle-consistent learning for single-channel unsupervised speech separation [11]. The cycle-consistency loss introduced in this study has been utilized in our proposal.

Let $\boldsymbol{x}_j$ be a mixture of two sound sources, $s_1^{(\boldsymbol{x}_j)}$ and $s_2^{(\boldsymbol{x}_j)}$. Suppose that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are given as mixtures of different source pairs, $s_1^{(\boldsymbol{x}_1)}$ and $s_2^{(\boldsymbol{x}_1)}$, and $s_1^{(\boldsymbol{x}_2)}$ and $s_2^{(\boldsymbol{x}_2)}$, which are all assumed to be different speech sources. First, these mixtures are segregated by the separator $G$ as follows:

$$\hat{\boldsymbol{s}}_1^{(\boldsymbol{x}_1)}, \hat{\boldsymbol{s}}_2^{(\boldsymbol{x}_1)} = G(\boldsymbol{x}_1), \quad \hat{\boldsymbol{s}}_1^{(\boldsymbol{x}_2)}, \hat{\boldsymbol{s}}_2^{(\boldsymbol{x}_2)} = G(\boldsymbol{x}_2), \quad (2)$$

where $\hat{\boldsymbol{s}}_i^{(\boldsymbol{x}_j)}$ denotes the $i$-th separated signal of the $j$-th observed mixture $\boldsymbol{x}_j$. These outputs are remixed in combinations between different observations as follows:

$$\boldsymbol{z}_1 = \hat{\boldsymbol{s}}_1^{(\boldsymbol{x}_1)} + \hat{\boldsymbol{s}}_2^{(\boldsymbol{x}_2)}, \quad \boldsymbol{z}_2 = \hat{\boldsymbol{s}}_1^{(\boldsymbol{x}_2)} + \hat{\boldsymbol{s}}_2^{(\boldsymbol{x}_1)}. \quad (3)$$

If speech separation in (2) works perfectly, the distribution of such *pseudo mixtures* will be consistent with that of the observed mixtures. $G$, therefore, can be trained adversarially with the discriminator $D$ that aims to discriminate between the observed mixtures (real) and the pseudo mixtures (fake). In addition to the loss function of this framework, $\mathcal{L}_{\mathrm{GAN}}$, two additional auxiliary loss functions are introduced as follows.

The pseudo mixtures $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are unmixed by the same separator $G$ as follows:

$$\hat{\boldsymbol{s}}_1^{(\boldsymbol{z}_1)}, \hat{\boldsymbol{s}}_2^{(\boldsymbol{z}_1)} = G(\boldsymbol{z}_1), \quad \hat{\boldsymbol{s}}_1^{(\boldsymbol{z}_2)}, \hat{\boldsymbol{s}}_2^{(\boldsymbol{z}_2)} = G(\boldsymbol{z}_2). \quad (4)$$

These outputs are remixed in the closest combination to the observed mixtures $\boldsymbol{x}_j$ into another pseudo mixtures as:

$$\hat{\boldsymbol{x}}_j = \left[\boldsymbol{A}\hat{\boldsymbol{S}}^{(\boldsymbol{z})}\right]_j, \quad \boldsymbol{A} = \min_{\boldsymbol{A}} \left\| \boldsymbol{x}_j - \left[\boldsymbol{A}\hat{\boldsymbol{S}}^{(\boldsymbol{z})}\right]_j \right\|, \quad (5)$$

where $\hat{\boldsymbol{S}}^{(\boldsymbol{z})} = \left[\hat{\boldsymbol{s}}_1^{(\boldsymbol{z}_1)}, \hat{\boldsymbol{s}}_2^{(\boldsymbol{z}_1)}, \hat{\boldsymbol{s}}_1^{(\boldsymbol{z}_2)}, \hat{\boldsymbol{s}}_2^{(\boldsymbol{z}_2)}\right]^\top \in \mathbb{C}^{2N\times M}$ denotes a set of separated signals of $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, and $\boldsymbol{A} \in \mathbb{B}^{2\times 2N}$ denotes a binary matrix that represents the assignment of each separated signal to $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$. The cycle-consistency loss (also referred to as remix-cycle-consistency loss in this paper) is defined as:

$$\mathcal{L}_C = \sum\nolimits_j ||\boldsymbol{x}_j - \hat{\boldsymbol{x}}_j||. \quad (6)$$

Since $\hat{\boldsymbol{s}}_i^{(\boldsymbol{x}_j)}$ contains residual noise and artifacts, the pseudo mixtures, $\boldsymbol{z}_j$, will contain unobserved distortions, which can make $\hat{\boldsymbol{s}}_i^{(\boldsymbol{z}_j)}$ distorted. It implies that distortion reduction in $\hat{\boldsymbol{s}}_i^{(\boldsymbol{x}_j)}$ leads to that in $\hat{\boldsymbol{s}}_i^{(\boldsymbol{z}_j)}$. Therefore, minimizing the loss $\mathcal{L}_C$, which is computed using $\hat{\boldsymbol{s}}_i^{(\boldsymbol{z}_j)}$, facilitates reducing the residual noise and artifacts in $\hat{\boldsymbol{s}}_i^{(\boldsymbol{x}_j)}$.

Here, it should be noted that a trivial solution $\hat{\boldsymbol{s}}_1^{(\boldsymbol{x}_j)} = \boldsymbol{x}_j$ and $\hat{\boldsymbol{s}}_2^{(\boldsymbol{x}_j)} = \boldsymbol{0}$ is unduly optimal both in $\mathcal{L}_{\mathrm{GAN}}$ and in $\mathcal{L}_C$. To compensate for this deficiency and to facilitate the separation, the following loss function is also introduced:

$$\mathcal{L}_E = \sum\nolimits_j \left( \left(\hat{\boldsymbol{s}}_1^{(\boldsymbol{x}_j)}\right)^2 + \left(\hat{\boldsymbol{s}}_2^{(\boldsymbol{x}_j)}\right)^2 \right). \quad (7)$$

$\mathcal{L}_C$ is designed to reduce distortions in the separated sound explicitly, but $\mathcal{L}_{\mathrm{GAN}}$ does not because it can be small even if $\hat{\boldsymbol{s}}_i^{(\boldsymbol{x}_j)}$ contains distortions depending on the combinations that make up $\boldsymbol{z}_j$. In fact, this method performed well in image separation but did not work in speech separation [11].

## 3. REMIX-CYCLE-CONSISTENT LEARNING ON ADVERSARIALLY-LEARNED SEPARATOR

This section describes an algorithm for training a separation network to explicitly remove residual noise and artifacts in the separated signals in the unsupervised setting. The overview of this method is illustrated in Fig. 1. The neural separator is adversarially trained to output a distortion-free speech-like signal coarsely and then tuned using a remix-cycle-consistency loss described in (6) to reduce distortions at the separator's output finely for each utterance. Note that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in Fig. 1 are different observed mixtures (i.e., they are not other channels of sound observed simultaneously).

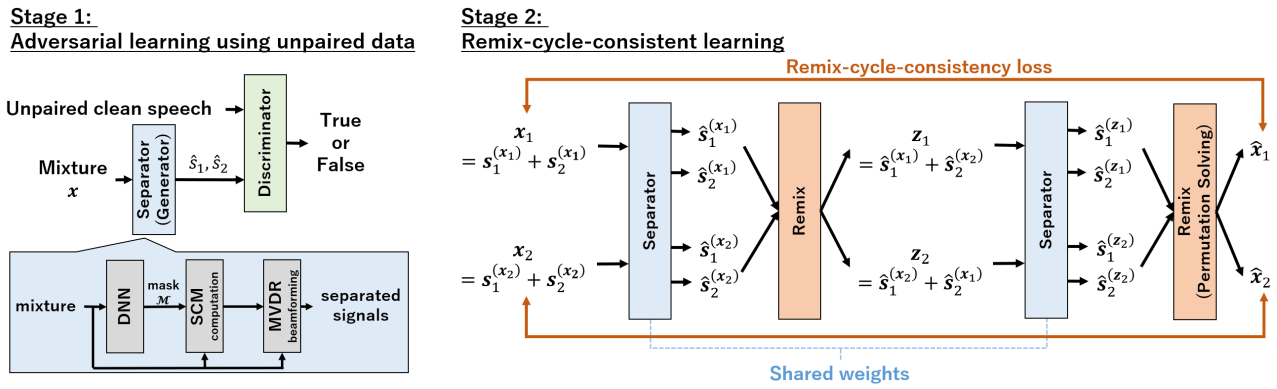


**Fig. 1**. Overview of proposed method. Neural separator (DNN-based MVDR beamformer) is adversarially trained and then tuned using remix-cycle-consistent learning. $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are different observed mixtures (not observed simultaneously).

### 3.1. Neural Separator

The separator employs a DNN-based MVDR beamformer [18–20], as illustrated at the bottom left of Fig. 1. The DNN takes a mixture as an input and estimates a TF mask of each source $\mathcal{M}_i$. The spatial covariance matrice (SCM) for the speech $\boldsymbol{R}_i^{(s)}$ and that for the noise $\boldsymbol{R}_i^{(n)} \in \mathbb{C}^{M\times M}$ are then computed using $\mathcal{M}_i$ and $\mathbf{1}-\mathcal{M}_i$, respectively [21]. The filter coefficient of the MVDR beamformer [22] is computed as:

$$\boldsymbol{W}_i = \frac{\boldsymbol{R}_i^{(n)}\boldsymbol{R}_i^{(s)}}{\mathrm{tr}(\boldsymbol{R}_i^{(n)}\boldsymbol{R}_i^{(s)})} \in \mathbb{C}^{M\times M}. \tag{8}$$

The $i$-th separated signal of a mixture $\boldsymbol{x}$ is given as:

$$\hat{\boldsymbol{s}}_i^{(\boldsymbol{x})} = \boldsymbol{W}_i^{\mathsf{H}}\boldsymbol{x} \in \mathbb{C}^{M}. \tag{9}$$

The pseudo mixtures need to be simulated as observed at the microphone locations to calculate the remix-cycle-consistency loss with the observed mixtures. To do this, the separator's output (i.e., estimated source) will be multiplied by the steering vector for the $M$-channel microphone positions. Note that (8) is formulated with such a process, and there is no need to apply the steering vectors in (9).

### 3.2. Learning Algorithm of Neural Separator

The neural separator, which corresponds to a generator, is adversarially trained with a discriminator that distinguishes the separator's outputs from unpaired clean speech. Since adversarial learning only aims to learn a distribution-to-distribution mapping, the performance of the separator trained with such a scheme can be limited and highly dependent on the initial values of the parameters.

The neural separator coarsely trained in the previous stage is then finely tuned using the remix-cycle-consistency loss introduced in [11]. As described in Sect. 2.2, this loss is designed to explicitly reduce the distortions in the separated signal for each utterance. Such a per-input-sample requirement is imposed on the learning instead of coarser constraints at the distribution level as in GANs, which can improve the separation performance and stabilize the learning.

The effectiveness of the proposed method will be further discussed in terms of 1) design for optimization and 2) design for loss functions. First, two-stage learning used in the proposed method is essential for improving the separation accuracy and stabilizing the learning. The remix-cycle-consistency loss may interfere with learning when the separator's performance is low because the trivial solution is optimal. This loss, therefore, should be used for tuning a well-trained separator, as in the proposed method. In contrast, the existing method [11] that minimizes all losses simultaneously did not work. Second, the adversarial loss in the proposed method is defined on the speech domain, i.e., to explicitly learn mixture-to-clean mapping, which is the aim of speech separation. In contrast, the adversarial loss in the existing method is defined on the mixture domain. In addition, the proposed method does not need the separation-promoting loss described in (7), unlike the existing method. The well-designed (i.e., not unduly flexible) neural separator, as well as two-stage learning, may help the proposed method avoid being trapped in the trivial solution.

## 4. SPEECH SEPARATION EXPERIMENT

Experimental comparisons were conducted to verify the effectiveness of the proposed method on the accuracy of speech separation and the stability of learning. To this end, the following learning methods were compared:

- **AL:** adversarial learning using unpaired clean speech (only the first stage in the proposed method)
- **AL + RCCL:** adversarial learning followed by remix-cycle-consistent learning (proposed method)
- **PIT:** permutation invariant training (upper limit)

Note that PIT [3] is a supervised learning method and will give an upper limit on AL and AL+RCCL, which are unsupervised learning methods. In addition, the adversarial unmix-and-remix is not included in the comparison because it was reported that it did not work in speech separation [11].

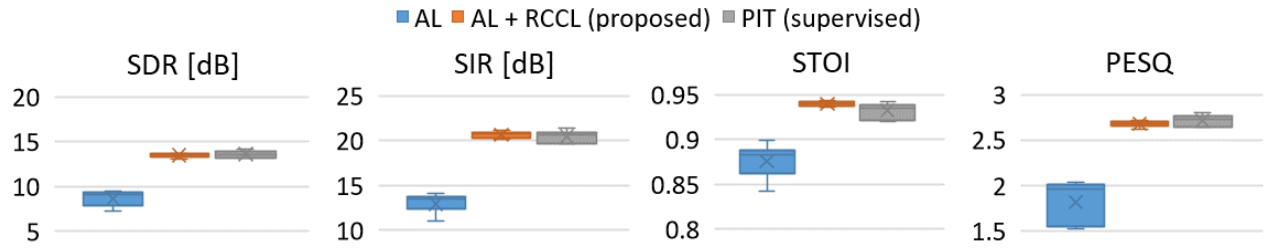


**Fig. 2**. Stability of adversarial learning (AL), AL followed by remix-cycle-consistent learning (AL+RCCL), and permutation invariant training (PIT), represented by box plots of SDR, SIR, STOI, and PESQ in ten trials (with different seeds) for learning.

**Table 1**. Speech separation performance of neural separator trained with adversarial learning (AL), AL followed by remix-cycle-consistent learning (RCCL), and permutation invariant training (PIT) in terms of SDR, SIR in decibels, STOI, and PESQ averaged over ten trials.

| Method | SDR | SIR | STOI | PESQ |
|---|---|---|---|---|
| Observation | 0.00 | 0.175 | 0.712 | 1.16 |
| USV: AL | 8.58 | 12.9 | 0.875 | 1.81 |
| USV: AL + RCCL | **13.4** | **20.6** | **0.939** | **2.68** |
| SV: PIT [3] | 13.6 | 20.4 | 0.932 | 2.72 |

### 4.1. Speech Materials

An anechoic sound field with two sources observed by four microphones whose spacings are 3 cm was simulated using pyroomacoustics [23]. The distance between the sound sources and the center of the microphone array was 1 m. The directions of two sound sources were randomly chosen without duplication between -90° and 90° at 15° intervals (0° is the front). Speech sources were sampled at 16 kHz and selected from the Wall Street Journal (WSJ0) corpus [24]. 10240, 1024, and 512 microphone observations (mixtures) were used for training, validation, and testing, respectively. Another 10240 distortion-free utterances were used as unpaired clean speech for adversarial learning. Speech signals were analyzed using the Hanning window with an FFT length of 512 and a hop size of 128.

### 4.2. Network Architecture

Neural networks were employed in the generator $G$ (specifically, the mask estimator in the neural separator) and the discriminator $D$. The mask estimator is composed of a fully connected layer with a rectified linear unit activation and two layers of bi-directional long short-term memory with 500 units followed by another fully connected layer with the softmax activation. $D$ consists of four two-dimensional convolutional layers followed by the sigmoid activation. All the parameters were optimized using the Adam optimizer [25]. In the first stage, the learning rate was set to $5.0 \times 10^{-4}$, and the batch size of $G$ and that of $D$ were set to 32 and 64, respectively. Note that each mixture consists of two speech sources, so the batch size of $D$ was twice that of $G$. In the second stage, the same learning rate was used, but the batch size was 16.

### 4.3. Experimental Results

The output signals of the neural separators trained with three learning methods and the microphone observation were evaluated using four criteria: the signal-to-distortion ratio (SDR), the signal-to-interference ratio (SIR), the short-time objective intelligibility (STOI) [26], and the perceptual evaluation of speech quality (PESQ) [27]. The average values of the performance over ten trials (i.e., the models created with ten different seeds) are listed in Table 1. The variation in performance, which represents the dependence on the initial values, is shown in Fig. 2 as box plots, where a five-number summary was the average value for 512 samples in the evaluation set.

Table 1 demonstrated that proposed two-stage learning (**AL+RCCL**) significantly improved the performance (e.g., about 5 dB in SDR) over the adversarial learning only (**AL**). Although the experiments were conducted in an anechoic condition, which is ideal for the proposed method, the performance of the proposed method is comparable to supervised learning, which shows the promise of the proposed method. Figure 2 clearly shows that **AL+RCCL** is much less sensitive to differences in seed values than **AL** in all metrics. For example, the distance in SDR between the minimum and maximum was more than 2 dB in **AL** while that in **AL+RCCL** was about 0.6 dB, which is even lower than that in **PIT** (i.e., 0.9 dB). These results suggest that the use of remix-cycle-consistent learning effectively increases the accuracy of speech separation and stabilizes adversarial learning.

## 5. CONCLUSION

This paper proposed a two-stage adversarial/remix-cycle-consistent learning for accurate and stable unsupervised speech separation. The proposed algorithm was designed to explicitly reduce the residual noise and artifacts in the output of the neural separator. The experimental results demonstrated that the proposed algorithm was effective in reducing the speech distortions and stabilizing adversarial learning.

## 6. REFERENCES


[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[2] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. ICASSP*, 2017, pp. 246–250.

[3] D. Yu, M. Kolbæk, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, 2017, pp. 241–245.

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

[5] P. Santiago, B. Antonio, and S. Joan, "Segan: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[6] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *Proc. ICASSP*, 2018, pp. 5039–5043.

[7] S. Fu, C. Liao, Y. Tsao, and S. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. ICML*, 2019.

[8] S. Jiaqi, J. Zeyu, and F. Adam, "Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *Proc. INTERSPEECH*, 2020.

[9] D. Chengyun, Z. Yi, M. Shiqian, S. Yongtao, S. Hui, and L. Xiangang, "Conv-tassan: Separative adversarial network based on conv-tasnet," in *Proc. Interspeech*, 2020, pp. 2647–2651.

[10] H. Phan, I. V. McLoughlin, L. Pham, O. Y. Chén, P. Koch, M. De Vos, and A. Mertins, "Improving gans for speech enhancement," *IEEE Signal Proc. Letters*, vol. 27, pp. 1700–1704, 2020.

[11] Y. Hoshen, "Towards unsupervised single-channel blind source separation using adversarial pair unmix-and-remix," in *ICASSP*, 2019, pp. 3272–3276.

[12] Y. Nakagome, M. Togami, T. Ogawa, and T. Kobayashi, "Efficient and stable adversarial learning using unpaired data for unsupervised multichannel speech separation," in *Proc. Interspeech*, 2021, pp. 3051–3055.

[13] T. Higuchi, K. Kinoshita, M. Delcroix, and T. Nakatani, "Adversarial training for data-driven speech enhancement without parallel corpus," in *Proc. ASRU*, 2017, pp. 40–47.

[14] M. Zhong, L. Jinyu, G. Yifan, and (Fred) J. Biing-Hwang, "Cycle-consistent speech enhancement," in *Proc. Interspeech*, 2018, pp. 1165–1169.

[15] J. Yuan and C. Bao, "Cyclegan-based speech enhancement for the unpaired training data," in *Proc. APSIPA ASC*, 2019, pp. 878–883.

[16] Y. Xiang and C. Bao, "A parallel-data-free speech enhancement method using multi-objective learning cycle-consistent generative adversarial network," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 28, pp. 1826–1838, 2020.

[17] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[18] H. Erdogan, J. Hershey, S. Watanabe, M. Mandel, and L. Jonathan, "Improved mvdr beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.

[19] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 196–200.

[20] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," in *Proc. ICML*, 2017, p. 2632–2641.

[21] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and H. Sawada, "A multichannel mmse-based framework for speech source separation and noise reduction," *IEEE Trans. on Audio, Speech, and Lang. Process.*, vol. 21, no. 9, pp. 1913–1928, 2013.

[22] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans on Audio, Speech, and Lang. Process.*, vol. 18, no. 2, pp. 260–276, 2010.

[23] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP*, 2018, pp. 351–355.

[24] J. S. Garofolo et al., *CSR-I (WSJ0) Complete LDC93S6A*, Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[25] P. K. Diederik and B. Jimmy, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[26] 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.

[27] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Process.*, 2001, vol. 2, pp. 749–752.