Permutation Invariant Training of Generative Adversarial Network for Monaural Speech Separation

Lianwu Chen[1†], *Meng Yu*[1†], *Yanmin Qian*[2‡], *Dan Su*[1], *Dong Yu*[3]

[1]Tencent AI Lab, Shenzhen, China
[2]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]Tencent AI Lab, Bellevue, WA, USA

{[1]lianwuchen, [1]raymondmyu, [1]dansu, [3]dyu}@tencent.com, [2]yanminqian@sjtu.edu.cn

Abstract

We explore generative adversarial networks (GANs) for speech separation, particularly with permutation invariant training (SSGAN-PIT). Prior work [1] demonstrates that GANs can be implemented for suppressing additive noise in noisy speech waveform and improving perceptual speech quality. In this work, we train GANs for speech separation which enhances multiple speech sources simultaneously with the permutation issue addressed by the utterance level PIT in the training of the generator network. We propose operating GANs on the power spectrum domain instead of waveforms to reduce computation. To better explore time dependencies, recurrent neural networks (RNNs) with long short-term memory (LSTM) are adopted for both generator and discriminator in this study. We evaluated SSGAN-PIT on the WSJ0 two-talker mixed speech separation task and found that SSGAN-PIT outperforms SSGAN without PIT and the neural networks based speech separation with or without PIT. The evaluation confirms the feasibility of the proposed model and training approach for efficient speech separation. The convergence behavior of permutation invariant training and adversarial training are analyzed.

Index Terms: permutation invariant training, generative adversarial networks, speech separation

1. Introduction

Human auditory system has a mechanism for separating mixed signals. Much research attention has been given to the topic of employing the machine to emulate human auditory perception. However, the progress made in multi-talker mixed speech separation, often referred to as the cocktail-party problem [2], has been less impressive. More generally, source separation is a relevant procedure in cases when a set of source signals of interest has gone through an unspecified mixing process and has been recorded at a single microphone or a microphone array. Given the observed mixture signal, the objective is to invert the unknown mixing process and estimate the individual source signals. Nevertheless, a truly general solution to source separation does not exist, e.g., the mixing mapping may be non-invertible.

Great advance was observed in monaural speech separation when the problem is converted into a supervised regression problem in which the optimization objective is closely related to the separation task. Inspired by the great success of deep learning on speech recognition, the deep learning based techniques have been developed to address the cocktail party problem recently [3]. These new techniques significantly outperformed the conventional approaches, such as minimum mean square error (MMSE) [4] suppressor, computational auditory scene analysis (CASA) [5, 6], and non-negative matrix factorization (NMF) [7, 8]. This framework and methodology works great for separating speech from noise and music, or speech of a specific known speaker from that of other speakers.

A recent breakthrough in the deep learning generative modeling field is generative adversarial networks (GANs) [9]. GANs have been successfully applied in the computer vision field to synthesize realistic images. The exploration of GANs on speech and audio has been limited. The speech enhancement GAN (SEGAN), proposed in [1] yields improvements to perceptual speech quality metrics over the noisy data and traditional enhancement baselines. This work has been further developed and evaluated by speech recognition in [10]. The generator network learns to model labeled data, e.g. the mapping from noisy speech samples to their clean counterparts, while the discriminator, usually a binary classifier, learns to discriminate between generated samples and target samples from training data. This framework is analogous to a two-player adversarial game, where minimax is a proven strategy. The key idea of GANs is to use the discriminator to shape the loss function of the generator.

In this work, we study the benefit of GANs for speech separation, where the generator network produces separated speech sources, while the discriminator tries to discriminate clean speech sources against enhanced speech sources. Unfortunately, the neural network based speech separation encounters the label ambiguity (or permutation) problem when applied to separate multiple speech streams from the mixed speech signal. As a result, the model cannot be effectively optimized and performs poorly on the cocktail party problem. Permutation invariant training (PIT) [11] is one of the most recent deep learning based approaches which achieved very impressive performance on addressing label permutation problem in speech separation. PIT casts speech separation as a multi-class segregation problem where the supervision is provided as a set instead of an ordered list. The permutation invariant training is integrated with the generator network, where the loss consists of an adversarial component and the distance between each generated speech source and its clean reference. Through the adversarial learning, it drives both models to improve their accuracy until generated samples are indistinguishable from real ones. To the best our knowledge, it is the first work on GANs for speech separation, particularly integrated with permutation invariant training.

The rest of the paper is organized as follows. In Section 2, the monaural speech separation is reviewed. In Section 3, we present details of our proposed generative adversarial networks with permutation invariant training for speech separation and

[†]Both authors contributed equally to this work.
[‡]Yanmin Qian did this work when he was a consultant in Tencent.

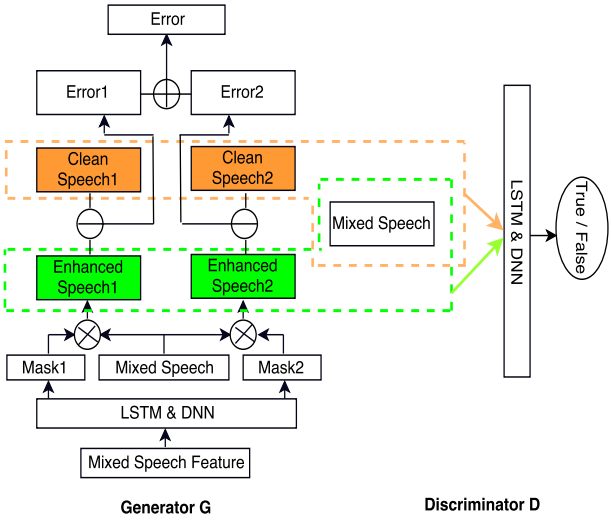

Figure 1: *Training diagram of SSGAN. Generator G maps mixed speech spectra to separated speech sources ("Enhanced Speech 1" and "Enhanced Speech 2"). D receives as input either concatenated components in the yellow or green box and decides if the triplet is real or enhanced*

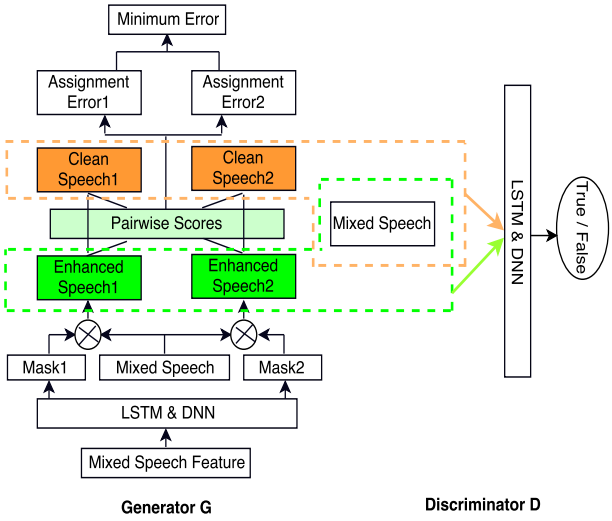

Figure 2: *Training diagram of SSGAN-PIT. The permutation invariant training is carried out in the training of generator G.*

discuss its advantages. We describe our experimental setup and evaluate the effectiveness of the proposed system in Section 4. We conclude this work in Section 5.

2. Monaural Speech Separation

The goal of monaural speech separation is to estimate the individual source signals from the mixture. Let us denote the S source signals in the time domain as $x_s(t), s = 1, \ldots, S$ and the microphone received mixed signal as $y(t) = \sum_{s=1}^{S} x_s(t)$. The corresponding spectrum representation by short-time Fourier transform (STFT) is $Y(t,f) = \sum_{s=1}^{S} X_s(t,f)$ for each time frame t and frequency subband f. Monaural speech separation is to recover each $X_s(t,f)$ from $Y(t,f)$. More specifically, we train a deep learning model $g(\cdot)$ such that $g(log|Y|^2;\theta) = \hat{M}_s, s = 1, \ldots, S$, where θ is the model parameter. We use log power spectrum for representing input noisy signal and the model infers idea ratio mask (IRM) $\hat{M}_s$. $\hat{M}_s(t,f) \geq 0$ and $\sum_{s=1}^{S} \hat{M}_s(t,f) = 1$ for all time-frequency bins (t,f). We then estimate source spectrogram $|X_s|$ as $|\hat{X}_s| = \hat{M}_s \otimes |Y|$, where $\otimes$ is the element-wise product of two operands. Due to the issue of zero-division in silence segments for label preparation, the cost function for regular deep learning based monaural speech separation is

$$\mathcal{J}_{\mathcal{SS}} = \frac{1}{T \times F \times S} \sum_{s=1}^{S} \| \hat{M}_s \otimes |Y| - |X_s| \|_F^2, \quad (1)$$

where $\| \cdot \|_F$ Frobenius norm. For simplicity and without lose of generality, we assume there are two-talkers in the signal mixture, i.e. $S = 2$ in the following discussion.

3. SSGAN-PIT

GANs consist of two components, a generator and a discriminator. The generator G maps latent vectors drawn from some known prior p_z to samples: $G : \mathbf{z} \rightarrow \hat{x}$, where $\mathbf{z} \sim p_z$. The way in which G learns to do the mapping is by means of an adversarial training, where we have another component, called the discriminator (D). D is in charge of transmitting information to G on what is real ($x \sim p_{data}$) and what is fake ($G(\mathbf{z}) \sim p_G$), such that G can slightly correct its output waveform towards the realistic distribution, getting rid of the noisy signals as those are signaled to be fake. In this sense, D can be understood as learning some sort of loss, for G's output to look real.

We propose to improve speech separation with GANs. In this model, the generator network G performs speech separation. It learns an effective mapping that can imitate the real speech distribution $\mathcal{X}$ to generate novel samples related to those of the training set. For speech separation, we used the conditional GANs, where the mixed speech signal Y is incorporated as conditional information in G and D. The outputs of the generator are the individual enhanced speech signals $|\hat{X}_s| = G(z, log|Y|^2), s = 1, \ldots, S$. Unlike speech denoising, the network G generates all the individual sources. G is designed to be an LSTM-RNN, similar to that used in the baseline monaural speech separation. D is an LSTM binary classifier whose input is either real samples, coming from the clean speech source dataset that G is imitating, or fake samples made up by G, i.e. enhanced individual sources. Specifically, the input to discriminator D is frame-wise concatenation of log power spectrum of the mixed signal and each individual clean (generated) signal, as shown in the yellow (green) box in Fig. 1, respectively.

It has been reported that the SEGAN generator learned to ignore the latent noise vector $\mathbf{z}$. According to [10], the latent vector is not assumed given the presence of noise in the input Y. We removed the latent vector from the generator. The training objective is simplified to be:

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{Y,X_s \sim p_{data}}[logD(|Y|, |X_{s=1:S}|)] + \mathbb{E}_{Y \sim p_{data}}[log(1 - D(|Y|, G(log|Y|^2)))] \quad (2)$$

G is trained to minimize this objective, while D is trained to maximize it.

In order to minimize the distance between the output of generator and the reference speech source $X_s, s = 1, \ldots, S$, we add a distortion term which leads to the new objective function

$$\min_G \max_D V(G,D) = \mathcal{J}_{SS} + \lambda \mathcal{L}_{cGAN}(G,D), \quad (3)$$

where $\mathcal{J}_{SS}$ is the monaural speech separation criterion defined in (1). This improved objective function encourages the adversarial component to generate more fine-grained and realistic samples. The importance of the two terms is balanced by a hyper-parameter λ. The whole training architecture is illustrated in Fig. 1. The generator on the left is based on a monaural speech separation model with an adversarial discrimination by the network D.

In order to stabilize training and increase the quality of generated samples in G, we refer to [1] to substitute the traditional GAN loss function with the least-square GAN objective. With this, the formulation for SSGAN in Eq. (3) changes to

$$\min_D V_{LSGAN}(D) = \frac{1}{2}\mathbb{E}_{Y,X_s\sim p_{data}}[(D(|Y|,|X_{s=1:S}|)-1)^2] + \frac{1}{2}\mathbb{E}_{Y\sim p_{data}}[D(|Y|,G(log|Y|^2))^2] \quad (4)$$

$$\min_G V_{LSGAN}(G) = \frac{\lambda}{2}\mathbb{E}_{Y\sim p_{data}}[(D(|Y|,G(log|Y|^2))-1)^2] + \mathcal{J}_{SS} \quad (5)$$

The generator G has multiple output layers, one for each mixing source, as shown in Fig. 1. Since both output layers depend on the same input mixture, reference assigning can be tricky especially if the training set contains many utterances spoken by many speakers. This problem is referred to as the label ambiguity (or permutation) problem. To address the label ambiguity problem, the permutation invariant training (PIT) was proposed in [11]. The architecture of PIT is shown as the left part of Fig. 2 which is integrated with generator G. In order to associate references to the output layers, the total number of $S!$ possible assignments between the references and the estimated sources are determined. We then compute the total MSE for each assignment, which is defined as the combined pairwise MSE between each reference $|X_s|$ and the estimated source $|\hat{X}_s|$. The assignment with the least total MSE is chosen and the model is optimized to reduce this particular MSE. In this work we adopt to use utterance-level Permutation Invariant Training (uPIT) [12], a simpler yet more effective approach to solve the tracing problem and the label permutation problem than original PIT. Specifically, we extend the frame-level PIT technique with the following utterance-level cost function:

$$\mathcal{J}_{\phi^*} = \frac{1}{T\times F\times S}\sum_{s=1}^{S} \parallel \hat{M}_s \otimes |Y| - |X_{\phi^*(s)}| \parallel_F^2, \quad (6)$$

where ϕ^* is the permutation that minimizes the utterance-level separation error defined as

$$\phi^* = \arg\min_{\phi\in\mathcal{P}} \sum_{s=1}^{S} \parallel \hat{M}_s \otimes |Y| - |X_{\phi(s)}| \parallel_F^2, \quad (7)$$

and $\mathcal{P}$ is the set of all $S!$ permutations. With uPIT, the permutation corresponding to the minimum utterance-level separation error is used for all frames in the utterance.

The permutation invariant training is implemented in SSGAN by updating the speech distortion term in the objective function (5) as below

$$\min_G V_{LSGAN}(G) = \frac{\lambda}{2}\mathbb{E}_{Y\sim p_{data}}[(D(|Y|,G(log|Y|^2))-1)^2] + \mathcal{J}_{\phi^*} \quad (8)$$

where $\mathcal{J}_{\phi^*}$ is the loss of utterance based PIT defined in (6). Fig. 2 describes the training strategy of SSGAN-PIT.

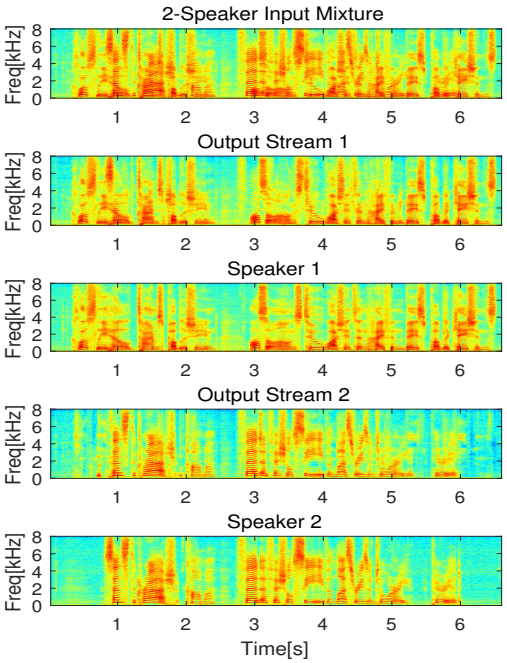

Figure 3: *Spectrograms showing SSGAN trained with PIT can separate a two-speaker mixture*

4. Experimental Results

4.1. Corpora

To study the capability of the proposed technique, we conducted experiments with mixtures of two speakers on the WSJ0-2mix dataset. The WSJ0-2mix dataset was introduced in [13] and was derived from the WSJ0 corpus [14]. The 30h training set and the 10h validation set contain two-speaker mixtures generated by randomly selecting from 49 male and 51 female speakers and utterances from the WSJ0 training set si_tr_s, and mixing them at various Signal-to-Noise Ratios (SNRs) uniformly chosen between 0 dB and 5 dB. The 5h test set was similarly generated using utterances from 16 unseen speakers from the WSJ0 validation set si_dt_05 and evaluation set si_et_05. We evaluated the baseline approaches and the proposed scheme on both close-condition test set (10h validation set) and open-condition test set (5h test set).

4.2. Architecture

The STFT of waveform signal is computed with a 32-ms window and shifted every 16ms. The generator G loads the 257 dimensional normalized log power spectrum, followed by 3 LSTM layers with 512 units per layer, one fully connected layer of 512 hidden units using rectified linear unit (ReLU) nonlinearity and a sigmoid output layer. Phase sensitive approximation [15] infers 257 dimensional real mask in the output layer.

The same network architecture is applied onto discriminator D except for the dimension of input feature layer and the output layer. As proposed in Section 3, we stack up normalized log power spectrum of mixed signal and individual sources as the input features for D. Therefore, the dimension of input layer is $257\times(S+1)$, where S is the number of mixed speech sources. Since two-source mixtures are employed in our evaluation, S equals to 2. The output label for D is 1 for real triplets (mixture-two clean sources) and 0 for fake (mixture-two

Table 1: *PESQ and SDR evaluation on WSJ0-2mix.*

Method	CC		OC	
	PESQ	SDR	PESQ	SDR
Original	1.86	2.09	1.89	2.12
Baseline SS	2.18	9.05	2.17	8.73
SSGAN	2.20	9.17	2.17	8.79
SS-PIT	2.40	11.14	2.39	10.86
SSGAN-PIT	2.44	11.26	2.41	10.90

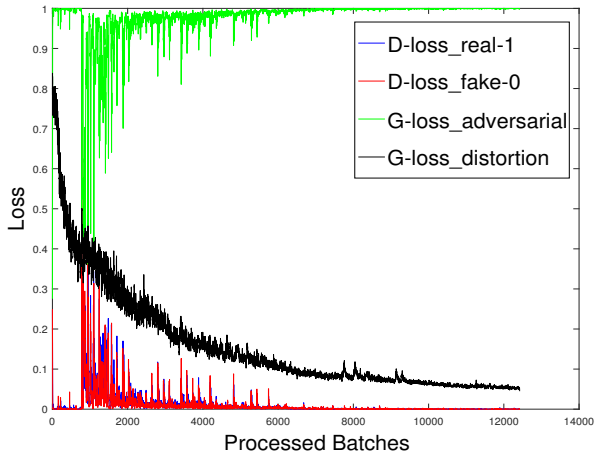

Figure 4: *The loss convergence of adversarial and distortion terms.*

enhanced sources) triplets.

4.3. Training

During GAN based training (SSGAN and SSGAN-PIT), for each mini-batch that contains 128 randomly selected utterance samples we alternate and update the D and G. To ensure the training converges to a good result efficiently, two different loss functions of G are used at two stages. For the first 5 epochs G is updated based on loss of speech distortion only, i.e. λ is set to 0. For the rest epochs, we set the importance weight to 0.1 which makes G keep correcting its output towards realistic generation. Results on different settings are also discussed in section 4.4. The whole network is trained for 80 epochs with Adam optimizer. The learning rate is initially set at 0.001 for training both generator and discriminator and then scaled down by 0.7 for every 20 epochs. A dropout with rate 0.2 is applied on each hidden layer in both G and D networks.

4.4. Results and Discussion

The SSGAN-PIT is evaluated on its potential to improve the Signal-to-Distortion Ratio (SDR) [16] and the Perceptual Evaluation of Speech Quality (PESQ) score [17], both of which are metrics widely used to evaluate speech separation tasks. As an example of SSGAN-PIT, Fig. 3 shows the spectrogram for a two-speaker (male-vs-female) test case.

In Table 1 we summarized the PESQ and SDR from different separation schemes for two-talker mixed speech in closed condition (CC) and open condition (OC). For a fair comparison, all approaches have been trained for 80 epochs. From Table 1 we can make several observations. First, SSGAN-PIT outperforms baseline monaural speech separation, its GAN based training and PIT based training, respectively. Second, with PIT, GAN based speech separation achieves a significant improvement when compared with that without PIT. This indicates that the permutation invariant training of spectral distortion is more critical than adversarial training, and furthermore the integration of the two achieves the best separation result. Third, the speech separation benefits from the adversarial training although the additional gain upon PIT is small. Fourth, same as the baseline approaches, SSGAN-PIT generalizes well on unseen speakers since the performances on the open and closed conditions are very close. We report the explorations on the input features to the discriminator D in Table 2. Compared with the concatenation of the mixture and individual sources ([mix, s_1, s_2]), two other variations are evaluated. All of them essentially perform similarly well. We observed that D converges faster on the proposed triplet ([mix, s_1, s_2]) than the two alternatives, probably due to the fact that such triplet encodes relationship between input components that is trivial to distinguish since $s_1 + s_2 = mix$ holds for real instances.

Table 2: *Evaluation of SSGAN-PIT by changing the inputs of D to pairs of source signals ($[s_1,\ s_2]$) and individual source signals ($[s_1]$ and $[s_2]$).*

Input of D	CC		OC	
	PESQ	SDR	PESQ	SDR
([mix, s1, s2])	2.44	11.26	2.41	10.90
([s1, s2])	2.43	11.22	2.40	10.80
([s1], [s2])	2.45	11.28	2.41	10.74

In Fig. 4 we present the SSGAN-PIT training progress as measured by the adversarial and spectral distortion loss, respectively, on the two-talker mixed speech training set WSJ0-2mix. The discriminator D converges to being perfectly discriminating "real" and "fake" clean speech samples as the loss on "real" (blue) and "fake" (red) samples converges to zero, respectively. In contrast, good performance of D raises the difficulty for the output of G to fool D, as indicated by the adversarial term in the loss of G (green). Nevertheless, the spectral distortion loss (black) is steadily decreasing as a function of batches, hence SSGAN-PIT, effectively separates the mixed speech sources.

5. Conclusions

In this paper, we proposed a novel scheme for monaural speech separation with adversarial and permutation invariant training. The experiment shows that the presented SSGAN-PIT outperforms those without adversarial and/or permutation invariant trainings. Although the evaluation was conducted for two-speaker mixture separation only, the presented training framework can handle mixtures with more than two speakers which will be studied in further experiments. In the context of speech separation, we explored to understand the behavior of the generator and discriminator through the convergence of both the adversarial loss and spectral distortion. As indicated in [10], training SEGAN with only the spectral distortion objective achieves better ASR performance than using the adversarial approach. As a next step, we will evaluate ASR performance on speech mixtures processed by SSGAN-PIT. Furthermore, we will investigate methods to balance G and D in the GAN training for the better results.

6. References

[1] S. Pascual, A. Bonafonte, and J. Serra, "Segan: speech enhancement generative adversarial network," in *Proc. INTERSPEECH*, 2017.

[2] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[3] Y. Qian, C. Weng, X. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 40–63, 2018.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[5] G. Brown and M. Cooke, "Computational auditory scene analysis." *Computer Speech & Language*, vol. 8, no. 4, pp. 297–336, 1994.

[6] D. Wang and G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[7] B. Schuller, F. Weninger, M. Wollmer, and et al., "Non-negative matrix factorization as noise-robust feature extractor for speech recognition," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2018, pp. 4562–4565.

[8] Z. Chen, B. McFee, and D. Ellis, "Speech enhancement by low-rank and convolutive dictionary spectrogram decomposition." in *Proc. INTERSPEECH*, 2014.

[9] 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 (NIPS), 2014*, 2014, pp. 2672–2680.

[10] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2018.

[11] D. Yu, M. Kolbak, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2017.

[12] M. Kolbk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[13] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2016, pp. 31–35.

[14] J. Garofolo, D. Graff, P. Doug, and D. Pallett, *CSR-I (WSJ0) Complete LDC93s6a*. Philadelphia: Linguistic Data Consortium, 1993.

[15] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2015, pp. 708–712.

[16] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[17] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 2001, pp. 749–752.