INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Conv-TasSAN: Separative Adversarial Network based on Conv-TasNet

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

Didi Chuxing, Beijing, China

{dengchengyun, yizhang, mashiqian, shayongtao, songhui, lixiangang}@didiglobal.com

Abstract

Conv-TasNet has showed competitive performance on single-channel speech source separation. In this paper, we investigate to further improve separation performance by optimizing the training mechanism with the same network structure. Motivated by the successful applications of generative adversarial networks (GANs) on speech enhancement tasks, we propose a novel Separative Adversarial Network called Conv-TasSAN, in which the separator is realized by using Conv-TasNet architecture. The discriminator is involved to optimize the separator with respect to specific speech objective metric. It makes the separator network capture the distribution information of speech sources more accurately, and also prevents over-smoothing problems. Experiments on WSJ0-2mix dataset confirm the superior performance of the proposed method over Conv-TasNet in terms of SI-SNR and PESQ improvement.
Index Terms: speech separation, separative adversarial networks, Conv-TasNet

1. Introduction

Speech is the most convenient and the fastest form of human communication. However, speech quality is often corrupted in ambient environments. We need to extract the desired speech from the complicated noisy background and keep the fidelity of the speech as much as possible. Therefore, speech separation technology is of great importance and has always been a research hotspot.

Speech separation algorithms can be roughly divided into two categories, time-frequency (T-F) methods and time-domain (or end-to-end) methods. T-F speech separation algorithms aim to estimate the enhanced spectrum of each individual source from the mixture spectra [1, 2, 3], and then rebuild the waveform via the inverse short-time Fourier Transform (iSTFT), by combining the enhanced magnitude spectrum with either noisy or the modified phase of the mixture. On the downside, the erroneous estimation of the phase limits the upper bound of separation performance [4].

To solve the phase problem, end-to-end speech separation methods such as the fully-convolutional time-domain audio separation network (Conv-TasNet) [4] and Wave-u-net [5] are proposed. Typically, Conv-TasNet comprises an encoder, a separator and a decoder. Several improvements have been proposed to Conv-TasNet recently. However, most of them concentrate on network architecture. For examples, the encoder/decoder based on a (deep) non-linear variant is proposed to replace the linear encoder/decoder in [6]. In [7], a deterministic gammatone filterbank is employed in encoder. In [8, 9]a parallel and multi-scale separator is proposed and in [10] a clustering mechanism is integrated into the separator. On the other hand, only a few works touch the training mechanism of deep-learning based speech separation. In [11], multi-task learning strategy and generative adversarial training are used to improve the performance of the CLDNN-based generator. [12] executes generative adversarial training (GAT) throughout the training to make the separated speech indistinguishable from the real one. These algorithms are both based on Least-square GAN (LSGAN) [13]. With fixed label training cascaded between two sections of Permutation Invariant Training (PIT) [1], [14] achieves better performance than PIT. In [15], a two-step training procedure is used to obtain better performance.

Generative adversarial networks (GANs) have not yet been well studied in speech separation, but they already have shown great potential in speech denoising tasks [16, 17]. Ordinary loss functions hold strong assumption on the distribution of the target, but GANs could overcome the limitation by its training mechanism [12]. To the best of our knowledge, Conv-TasNet has not been extended to use generative adversarial training to improve the performance.

In this study, we incorporate Conv-TasNet into a GAN-based speech separation system. Instead of using generative adversarial network, we propose to use separative adversarial network, which is labeled as Conv-TasSAN in this paper. Conv-TasSAN consists of a Conv-TasNet separator and an encoder-TCN(temporal convolutional network)-CNN discriminator. Our contributions mainly focus on two fields: utterance-level permutation invariant training (uPIT) [2] based separative adversarial network architecture and using speech objective metrics as discriminator targets. The second contribution effectively overcomes the discriminator-evaluation mismatch (DEM), and leads the separator to learn the source distribution better. We will further show that Conv-TasSAN leads to a scale-invariant source-to-noise ratio (SI-SNR) improvement of 0.7dB and a perceptual evaluation of speech quality (PESQ) [18] improvement of 0.1 in WSJ0-2mix test dataset [19] comparing with Conv-TasNet, while the separator has the same parameters setting as Conv-TasNet.

We will first give an overview on our experimental framework in Section 2 and then go into details about the proposed Conv-TasSAN in Section 3. In Section 4 we present our results and then come to the conclusions in Section 5.

2. Relation to Prior Work

Single-channel multi-speaker speech separation can be formulated in terms of extracting C sources $s_1(t)$, ..., $s_C(t) \in \mathbb{R}^{1\times T}$ from the mixture $x(t) \in \mathbb{R}^{1\times T}$ (T is the segment of the mixture) , where

$$x(t) = \sum_{i=1}^{C} s_i(t) \tag{1}$$

http://dx.doi.org/10.21437/Interspeech.2020-2371

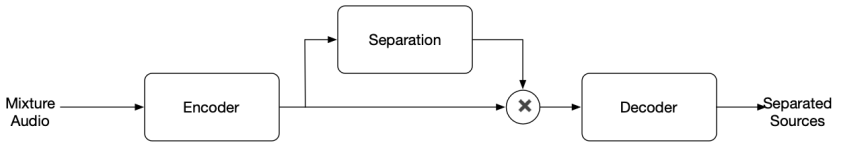


Figure 1: *The block diagram of the Conv-TasNet system.*

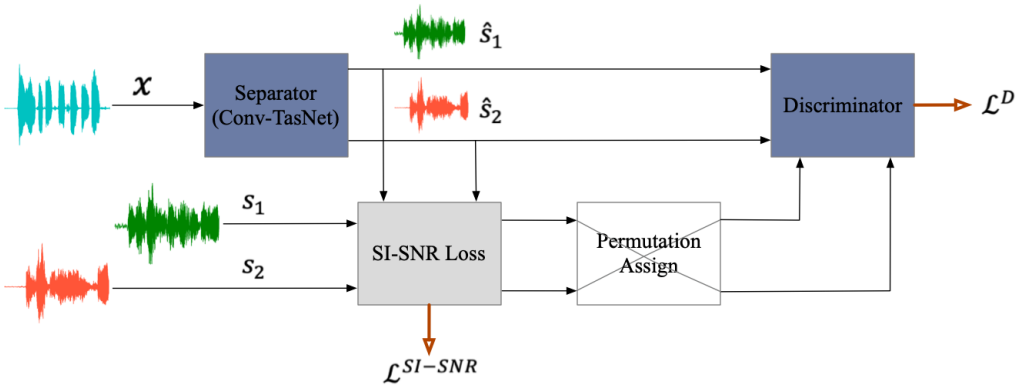


Figure 2: *Architecture of the separator and discriminator in Conv-TasSAN. Sources are separated by the separator Conv-TasNet and separated speech are processed by a discriminator together with ground-truth sources. During training, the clean permutation problem of the discriminator is solved based on SI-SNR calculation.* $\pounds^{SI\text{-}SNR}$ *and* $\pounds^{D}$ *are the loss of SI-SNR and D.*

2.1. Conv-TasNet

Conv-TasNet is a fully convolutional time-domain speech separation network, with encoder, decoder and separation block, as illustrated in Figure 1. The input mixture speech can be divided into K overlapping segments with length L and stride $L/2$, which is represented as matrix $\mathbf{x} \in \mathbb{R}^{K \times L}$. And then $\mathbf{x}$ is transformed to a latent space with $K \times N$ dimension by a 1-D convolution operation. That is, output of the encoder can be defined as,

$$E = ReLU(\mathbf{x}U) \tag{2}$$

where $U \in \mathbb{R}^{L \times N}$ is the encoder basis functions matrix, N is the kernel size of encoder, $ReLU$ is the rectified linear unit to ensure that the representation is non-negative.

The separation block is a temporal convolutional network (TCN) [20]. TCN consists of multiple stacked 1-D dilated convolutional blocks with increasing dilation factors, and predicts a representation for each source by learning a mask $m_i \in \mathbb{R}^{K \times N}$ in this latent space. The representation of the i^{th} source $\hat{d}_i$ can be calculated by applying the corresponding mask to encoder output of the mixture,

$$\hat{d}_i = E \odot m_i \tag{3}$$

where $\odot$ denotes element-wise multiplication.

Finally, the overlapping-segment waveform of the i^{th} source $\hat{\mathbf{s}}_i \in \mathbb{R}^{K \times L}$ is reconstructed by the decoder consisted of 1-D transposed convolution,

$$\hat{\mathbf{s}}_i = \hat{d}_i V \tag{4}$$

where $V \in \mathbb{R}^{N \times L}$ is the decoder basis functions matrix. The overlapping reconstructed matrix generates the final waveform $\hat{s}_i \in \mathbb{R}^{1 \times T}$ by overlap and sum operation.

The training objective of Conv-TasNet is maximizing the scale-invariant source-to-noise ratio (SI-SNR) [21] with utterance-level permutation invariant training. SI-SNR is defined as:

$$\begin{cases} s_{target} := \frac{<\hat{s}, s> s}{\|s\|^2} \\ e_{noise} := \hat{s} - s_{target} \\ SI\text{-}SNR := 10log_{10} \frac{\|s_{target}\|^2}{\|e_{noise}\|^2} \end{cases} \tag{5}$$

where $\hat{s}$ and s are the predicted and original clean sources, respectively, and $\|s\|^2 =< s, s >$ denotes the signal power. SI-SNR loss of Conv-TasNet is a conventional loss function, and it's strong assumption on how the distribution of the estimated sources is shaped introduces an upper bound on the separation performance.

2.2. MetricGAN

GAN has been an active research area in recent years, because of the capability of modeling a complex data distribution of ground-truth target. A function $Q(I)$ (where I is the input of the metric) is introduced in the discriminator of MetricGAN [17] to solve the discriminator-evaluation mismatch (DEM) problem on speech enhancement, which represents the evaluation metric. For speech enhancement task, $Q(I)$ are PESQ and STOI [22] scores of the estimated speech and the corresponding clean speech.

In MetricGAN, the generator (G) generates time-frequency masks (between 0 to 1) and the masks are multiplied with noisy magnitude spectrogram. The enhanced spectrogram together with the corresponding clean spectrogram are fed into the discriminator (D) to get the estimated metric scores as feedback to G. That is, G needs to generate enhanced speech with higher metric scores and D needs to behave similar to Q. Loss of G

and D are defined as,

$$
\begin{aligned}
min\mathcal{L}(D) = \mathbb{E}_{x,y}[(D(y,y)-1)^2] + \\
(D(G(x),y) - Q^{'}(G(x),y))^2] \\
min\mathcal{L}(G) = \mathbb{E}_{x}[(D(G(x),y)-1)^2]
\end{aligned} \tag{6}
$$

where $\mathbb{E}$ is the expectation operator, $Q^{'}$ stands for the normalized evaluation metric, x is the noisy speech and y is the corresponding clean speech.

3. Separative Adversarial Network based on Conv-TasNet

Discriminator of GAN provides a high-level abstract measurement of realness [23], which can release the over-smoothing problems in the loss minimizing process caused by the strong assumption of conventional losses. Motivated by MetricGAN, we execute separative adversarial training to Conv-TasNet, labeled as Conv-TasSAN. The proposed approach includes two major components: Conv-TasSAN architecture consisting of a Conv-TasNet separator and TCN-based discriminator; and the novel loss function.

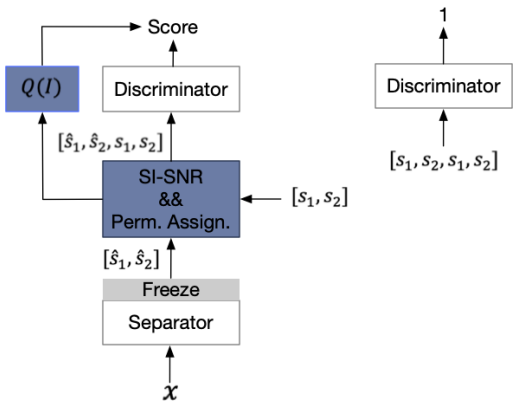
In this section, we first give the details of the Conv-TasSAN architecture. And then, we introduce the perceptual metric loss function and separative adversarial training mechanism.

3.1. Network architecture

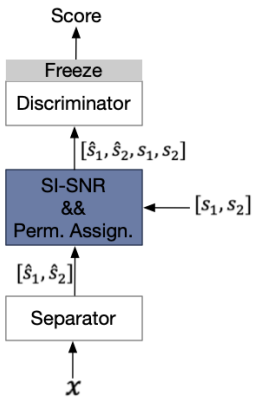
The Conv-TasSAN separation system comprises a separator and a discriminator, and the architecture is illustrated in Figure 2. The separator is Conv-TasNet [4], which takes the mixture waveform as the input and estimates the separated waveform of each source. Separation outputs are re-aligned to the clean source order, according to the maximum SI-SNR criterion or PIT. Aligned separation output and clean sources are then contacted as the input of the discriminator. Discriminator evaluates the degree of realness as [17], instead of just distinguishing real and fake. That is, the output of discriminator is the metric score between 0 and 1 (1 represents the best score of clean source).

Our implementation of the separator consists of the same architecture and parameter configuration as described in [4]. The encoder is a 1-D convolutional layer. The number of filters is 512, the kernel size is 16 in samples and the stride is 8. The separation block is a TCN, which consists of B stacked 1-D convolutional blocks and each stack is repeated R times. In our separator, we keep $B = 8$ and $R = 3$ same as in [4]. Number of filters in depthwise convolutional layers of TCN is 512. The decoder is a transposed convolutional layer of the encoder.

The discriminator has a similar structure to the separator except decoder, which consists of an encoder and a TCN. The encoder is a 1-D convolutional layer. The number of filter is 256, the kernel size is 16 and the stride is 8. In our discriminator, we use $B = 8$ and $R = 2$ for the TCN. Number of filters in depthwise convolutional layers of TCN is 256. It is not necessary to model the details of sources as in the separator. Two 1-D convolutional layers are following the TCN. The number of filters in the respective convolutional layers are set to 8 and 1. The kernel size is 15 for the first layer and 1 for the second layer. Finally, a fully connected layer with 1 linear node is added subsequently. In the discriminator, we replace PReLU with LeakyReLU, and use global layer normalization in all hidden layers.



(a) Step 1: Update D



(b) Step 2: Update S

Figure 3: *Training procedure for Conv-TasSAN. Discriminator D and Separator S are trained alternatively.* $\hat{s_1}$ *and* $\hat{s_2}$ *are the separated sources estimated by the separator,* s_1 *and* s_2 *are the clean sources. Function* $Q(I)$ *represents the evaluation metric to be optimized, where* I *is the input of the metric.*

Table 1: *SI-SNRi and SDRi in dB for the test dataset of the WSJ0-2mix database.*

Algorithms	Params.	SI-SNRi	SDRi
Conv-TasNet [4]	5.1M	15.3	15.6
Conv-TasNet (ours)	5.0M	14.4	14.7

3.2. Separative adversarial training and loss function

In this work, the separator is trained via separative adversarial training (SAT). We take two-speaker separation task for example. The training procedure is shown in Figure 3. Discriminator and separator are trained alternatively in each individual epoch:

Step 1 - Discriminator training:

Train the discriminator to minimize the distance from $Q(I)$ between separation output and clean sources. First of all, separation output $\hat{s_1}$ and $\hat{s_2}$ are estimated by the separator, and then SI-SNR scores are calculated on each permutation combination of separation output and clean source. Secondly, reorder the separation output according to the maximum SI-SNR. The out-of-order inputs introduce information redundancy into the discriminator resulting in severe performance degradation, which has been verified by our experiments. That is, the discriminator need additional parameters to learn the ordering of signal

Table 2: *Comparison of performances of different separation methods on the WSJ0-2mix dataset. Best scores are highlighted in* ***bold****.*

Algorithms	Params. of S	Params. of D	Target of D	SI-SNRi	SDRi	PESQi	STOIi
CBLDNN-GAT [11]	39.5M	-	Fake/True	-	11.0	-	-
FurcaX [12]	-	-	Fake/True	-	12.5	-	-
Conv-TasNet (ours)	5.0M	-	-	14.4	14.7	1.07	0.20
Conv-TasSAN	5.0M	1.3M	STOI	14.8	15.1	1.13	0.21
Conv-TasSAN	5.0M	1.3M	PESQ	**15.1**	**15.4**	**1.17**	**0.22**

combinations. Then, train the discriminator to have similar behavior as metric Q, that is, get accurate scores on both separated data and clean data. Most speech distortion metrics are suitable for the discriminator, such as PESQ and STOI. In this paper, we use utterance-level PESQ or STOI as the metric target for more reasonable direction to connect the model training with the goal of speech enhancement. To avoid negative Q_{PESQ} and get smoother training, we map Q_{PESQ} to range $[0, 1]$ by,

$$Q'_{PESQ} = (Q_{PESQ} + 0.5)/5.0 \tag{7}$$

To ensure the discriminator behaves similarly to Q, we simply modify the objective function of discriminator,

$$\begin{aligned} min\pounds(D) = \mathbb{E}[((D(S(x), s_1, s_2) - Q(S(x), s_1, s_2))^2 \\ +(D(s_1, s_2, s_1, s_2) - 1)^2] \end{aligned} \tag{8}$$

where $S(x)$ stands for the output of the separator, that is, $\hat{s_1}$ and $\hat{s_2}$. In particular, for some extreme training examples where PESQ scores cannot be calculated, we use $1e-5$ as their PESQ scores rather than 0.

By associating the discriminator with PESQ or STOI and PIT, the discriminator can solve the problem that the lower scores of objective metrics obtained by the time-domain model comparing with time-frequency model on separation task. Wherefore, Conv-TasSAN can capture the behavior of the PESQ or STOI and provide accurate gradients guiding the separator updates, comparing with using SI-SNR loss only.

Step 2 – Separator training:

Train the separator according to the gradient provided by the discriminator. The separator is trained to make the discriminator output a score close to 1. Meanwhile, uSI-SNR loss is used as a regularization to guide the training,

$$\begin{aligned} min\pounds(S) = \lambda\mathbb{E}[((D(S(x), s_1, s_2) - 1)^2 + \\ uSI\text{-}SNR(S(y), s_1, s_2) \end{aligned} \tag{9}$$

where λ is a hyperparameter used to balance SAT and $uSI\text{-}SNR$ loss.

4. Experimental Evaluation

4.1. Dataset

We evaluate our system on two-speaker speech separation problem using WSJ0-2mix dataset [19], which contains 30 hours of training data and 10 hours of validation data. The mixtures are generated by randomly selecting 49 male and 51 female speakers and utterances in Wall Street Journal (WSJ0) training set si_tr_s, and mixing them at various signal-to-noise ratios (SNR) uniformly between 0 dB and 5 dB. 5 hours of evaluation set is generated in the same way, using utterances from 16 unseen speakers from si_et_05 in the WSJ0 dataset. To reduce the computational cost, all of the waveforms are down-sampled to 8 kHz.

4.2. Training and evaluation setup

All separators including the original Conv-TasNet are trained using Adam optimizer [24] for 100 epochs with an initial learning rate of 0.001 using a batchsize of 8, and the learning rate is divided by 2 if the validation loss does not improve for 2 epochs. We use Adam with a fixed learning rate of 0.0005 on the discriminator. All parameters of models are initialized by xavier normal on Pytorch. The hyper-parameters that control the discriminator loss of the separator is set as $\lambda = 10$ such that it has the same order of magnitude with respect to the SI-SNR loss.

We use the SI-SNR improvement (SI-SNRi), signal-to-distortion ratio improvement (SDRi) [25], PESQ improvement (PESQi) and STOI improvement (STOIi) as objective measures of separation accuracy.

4.3. Results

We implement the Conv-TasNet by following the hyper-parameter notations in the original paper [1] . SI-SNRi and SDRi are listed in Table 1. It can be seen that our implementation yields slight degradation in performance on the test dataset of the WSJ0-2mix database comparing with [4], which might be caused by different initialization of training and lack of clip norm.

The results for the different algorithms are show in Table 2. We notice that the proposed Conv-TasSAN algorithm leads to a consistent improvement over the Conv-TasNet implemented by us where we train the same architecture of separator, whether the discriminator with targets PESQ or STOI. Conv-TasSAN with PESQ outperforms Conv-TasSAN with STOI. This may be attributed to the fact that PESQ has greater dynamic range than STOI on the same dataset. Our Conv-TasSAN with PESQ metric yields an absolute SI-SNR improvement over the baseline of up to 0.7 dB, an absolute PESQ improvement of 0.10. Thus, Conv-TasSAN with PESQ can further improve the comprehensive performance comparing with the Conv-TasNet.

5. Conclusions

This paper introduces a novel separative adversarial network based on Conv-TasNet called Conv-TasSAN. We use PESQ or STOI as the discriminator target to improve the accuracy of sources distribution modeled by the separator. The proposed method significantly improves the source separation performance over the baseline framework. As future works, we shall explore more separative adversarial networks based on more GANs and more speech separation architectures.

[1]Our implement is based on https://github.com/JusperLee/Dual-Path-RNN-Pytorch. We add skip-connection to the original code, remove bias in all layers and change the 1-D (transposed) convolutional layer of the decoder to fully-connected layer together with overlap-and-sum according to [4]. Clip norm is not using in our experiment.

6. References

[1] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[2] M. Kolbæk, 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.

[3] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 686–690.

[4] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[5] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[6] B. Kadioglu, M. Horgan, X. Liu, J. Pons, D. Darcy, and V. Kumar, "An empirical study of conv-tasnet," *arXiv preprint arXiv:2002.08688*, 2020.

[7] D. Ditter and T. Gerkmann, "A multi-phase gammatone filterbank for speech separation via tasnet," *arXiv preprint arXiv:1910.11615*, 2019.

[8] Z. Shi, H. Lin, L. Liu, R. Liu, J. Han, and A. Shi, "Deep attention gated dilated temporal convolutional networks with intra-parallel convolutional modules for end-to-end monaural speech separation," in *Proc. Interspeech*, 2019, pp. 3183–3187.

[9] Z. Shi, H. Lin, L. Liu, R. Liu, S. Hayakawa, S. Harada, and J. Han, "End-to-end monaural speech separation with multi-scale dynamic weighted gated dilated convolutional pyramid network," in *Proc. Interspeech*, 2019, pp. 4614–4618.

[10] G.-P. Yang, C.-I. Tuan, H.-Y. Lee, and L.-s. Lee, "Improved speech separation with time-and-frequency cross-domain joint embedding and clustering," *arXiv preprint arXiv:1904.07845*, 2019.

[11] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, "Cbldnn-based speaker-independent speech separation via generative adversarial training," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 711–715.

[12] Z. Shi, H. Lin, L. Liu, R. Liu, S. Hayakawa, and J. Han, "Furcax: End-to-end monaural speech separation based on deep gated (de) convolutional neural networks with adversarial example training," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6985–6989.

[13] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[14] G.-P. Yang, S.-L. Wu, Y.-W. Mao, H.-y. Lee, and L.-s. Lee, "Interrupted and cascaded permutation invariant training for speech separation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6369–6373.

[15] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, "Two-step sound source separation: Training on learned latent targets," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 31–35.

[16] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 106–110.

[17] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," *arXiv preprint arXiv:1905.04874*, 2019.

[18] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

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

[20] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 47–54.

[21] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[22] 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.

[23] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," *arXiv preprint arXiv:1807.07501*, 2018.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.