INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

UnSE: Unsupervised Speech Enhancement Using Optimal Transport

Wenbin Jiang[1*], *Fei Wen*[2], *Yifan Zhang*[1], *Kai Yu*[1*]

[1]X-LANCE Lab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China
{jwb361,wenfei,zhang_yifan,kai.yu}@sjtu.edu.cn

Abstract

Most deep learning-based speech enhancement methods usually use supervised learning, which requires massive noisy-to-clean training pairs. However, the synthesized training data can only partially cover some realistic environments, and it is generally difficult or almost impossible to collect pairs of noisy and ground-truth clean speech in some scenarios. To address this problem, we propose an unsupervised speech enhancement method that does not require any paired noisy-to-clean training data. Specifically, based on the optimal transport criterion, the speech enhancement model is trained in an unsupervised manner only using a noisy speech based fidelity loss and a distribution divergence loss, by which the divergence between the output and (unpaired) clean speech is minimized. Experimental results show that the proposed unsupervised method can achieve competitive performance with supervised methods on the VCTK + DEMAND benchmark and better performance on the CHiME4 benchmark.

Index Terms: Unsupervised learning, speech enhancement, optimal transport, deep learning, generative adversarial networks

1. Introduction

Speech enhancement aims to estimate the unseen clean speech from the observed noisy speech to improve the perceptual speech quality or recognition accuracy. The traditional speech enhancement algorithms, based on the Wiener filter or statistical model, have been studied for decades [1]. In recent years, thanks to the development of deep learning technology [2, 3], speech enhancement methods based on deep neural networks (DNNs) have made much progress [4, 5, 6, 7, 8], with denoising performance significantly better than traditional methods.

Typically, DNN-based speech enhancement models are trained in a supervised manner [9]. The training procedure requires pairs of noisy speech as the inputs and ground-truth clean speech as the outputs, and the noisy speech utterance is simulated by adding noise to the clean speech. In order to improve the generalization performance of the denoising model, massive (e.g., thousands of hours of) noisy-to-clean training data are usually used. However, the simulated noisy speech cannot cover all the realistic environments. Once a domain mismatch exists between simulation and reality, the performance of the denoising model would deteriorate significantly. Moreover, collecting pairs of noisy speech and ground-truth clean speech in some practical application scenarios is generally difficult or almost impossible. This motivates us to develop an unsupervised learning method for speech enhancement that does not require any paired noisy-to-clean data for training.

Several semi-supervised and self-supervised learning methods have been proposed for speech enhancement to reduce the reliance on paired training data. In [10], a RemixIT training scheme, which does not need isolated in-domain speech and noise, is proposed to train the speech enhancement model only using noisy speech. In [11], a denoising autoencoder with a linear regression decoder framework is proposed for speech enhancement. The model is trained in a self-supervised learning fashion and uses noisy speech as both the input and the training target. However, the RemixIT is trained with a teacher-student training framework that requires a pre-trained teacher model, and the work [11] requires setting empirical parameters in calculating the linear transformation matrix. On the other hand, several unsupervised adaptation methods based on optimal transport have been proposed for speech enhancement [12] and spoken language identification [13]. The work [12] proposes a discriminator-constrained optimal transport network (DOTN) for exploiting the domain knowledge. The work [13] uses a similar neural adaptation model to address the statistical distribution mismatch between the training and testing data. The results in [12, 13] show that introducing an additional optimal transport neural network makes minimizing the speech signal distribution between the source and target domains feasible.

Inspired by the unsupervised noise adaptation for speech enhancement [12] and our recent work of optimal transport for unsupervised image denoising [14], this paper proposes a novel unsupervised speech enhancement method without the need for paired noisy-to-clean training data. Concretely, based on the optimal transport theory, the denoising learning problem can be formulated as unsupervised learning with a constraint on the output distribution and can be further relaxed into an unconstrained optimization problem for implementation. The speech enhancement model is trained in an unsupervised manner only using a noisy speech based fidelity loss and a distribution divergence loss. The discriminator is optimized to minimize the probability distribution between the estimated and unpaired clean speech. To the best of our knowledge, we are the first to use optimal transport for unsupervised speech enhancement. The main contributions of this work are as follows: *1)* We propose a novel unsupervised speech enhancement learning method that does not require any paired noisy-to-clean training data but achieves competitive performance with supervised methods. *2)* We conducted extensive speech recognition experiments and found that the proposed method is superior to the supervised method, especially using the test time adaptation.

2. Optimal Transport for Unsupervised Denoising Learning

Let $\mathbb{P}_\mathcal{Y}$ and $\mathbb{P}_\mathcal{X}$ denote the probability distributions of two datasets $\mathcal{Y}$ and $\mathcal{X}$, respectively. Optimal transport aims to find

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62271314.

*corresponding author.

10.21437/Interspeech.2023-378

the most efficient transport from $\nu \sim \mathbb{P}_{\mathcal{Y}}$ to $\mu \sim \mathbb{P}_{\mathcal{X}}$ that minimizes the total cost [15]

$$\begin{aligned} &\inf_f \int_{\mathcal{Y}} c(f(y), y) d\nu(y) \\ &s.t. \quad f_{\#}\nu = \mu, \end{aligned} \tag{1}$$

where $f : \mathcal{Y} \to \mathcal{X}$ is a transport map that transports $\nu \sim \mathbb{P}_{\mathcal{Y}}$ to $\mu \sim \mathbb{P}_{\mathcal{X}}$, $c : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^+$ is the cost function of the transportation, $f_{\#}\nu$ denotes the transport of ν by f, and $f_{\#}\nu = \mu$ is the constraint on the transported distribution.

In the absence of paired noisy-to-clean training data, the unsupervised denoising learning using optimal transport can be formulated as [14]

$$\begin{aligned} &\min_f \mathbb{E}_{y \sim \mathbb{P}_{\mathcal{Y}}} (||y - f(y)||_p^p) \\ &s.t. \quad \mathbb{P}_{f(\mathcal{Y})} = \mathbb{P}_{\mathcal{X}}, \end{aligned} \tag{2}$$

where $f(\cdot)$ is the denoising neural network, and $|| \cdot ||_p$ is the ℓ_p-norm with $p \geq 1$. Obviously, the optimization problem (2) is an application of the optimal transport problem in (1).

In implementation, the constrained optimization problem in (2) can be relaxed to an unconstrained one as

$$\min_f \mathbb{E}_{y \sim \mathbb{P}_{\mathcal{Y}}} (||y - f(y)||_p^p) + \lambda d(\mathbb{P}_{f(\mathcal{Y})}, \mathbb{P}_{\mathcal{X}}) \tag{3}$$

where $d(\cdot, \cdot)$ measures the divergence between two probability distributions, and $\lambda > 0$ is a balance factor. It has been proved that, the formula in (3) has the same solution as the constrained one in (2) under certain conditions [14].

3. The Proposed Method

3.1. Problem Formulation

For speech signal, we operate on the short-time Fourier transform (STFT) domain and consider the additive signal model,

$$\mathbf{Y} = \mathbf{X} + \mathbf{N}, \tag{4}$$

where $\mathbf{Y} \in \mathbb{C}^{F \times T}$, $\mathbf{X} \in \mathbb{C}^{F \times T}$, and $\mathbf{N} \in \mathbb{C}^{F \times T}$ are the STFT of the time domain noisy speech $\mathbf{y}$, clean speech $\mathbf{x}$, and additive noise $\mathbf{n}$, respectively. F and T denote the number of frequency bins and frames, respectively.

The goal of speech enhancement is to find a nonlinear function (i.e., a denoising neural network) f_θ such that $\hat{\mathbf{X}} = f_\theta(\mathbf{Y}) \approx \mathbf{X}$. The most popular supervised learning method is training the denoising network f_θ that minimizes the loss of the denoised speech and clean speech, i.e., $\mathcal{L}(f_\theta(\mathbf{Y}), \mathbf{X})$. In contrast, we propose minimizing the loss of the denoised and observed noisy speech, i.e., $\mathcal{L}(f_\theta(\mathbf{Y}), \mathbf{Y})$. More specifically, the denoising network is learning in a purely unsupervised manner. To prevent the denoising network from just learning an identity function, we use an additional adversarial loss of the unpaired clean speech, which is inspired by our recent work of optimal transport for unsupervised image denoising [14].

3.2. Unsupervised Speech Enhancement

Figure 1 illustrates the diagram of the proposed unsupervised speech enhancement training method. The noisy speech $\mathbf{y} \sim \mathbb{P}_\mathbf{y}$ and clean speech $\mathbf{x} \sim \mathbb{P}_\mathbf{x}$ are independent, i.e., they are unpaired training data. f_θ denotes the generator (i.e., the denoising model), f_ϕ denotes the discriminator, and they are trained in an adversarial training process, using generative adversarial networks (GANs) [16]. $\mathcal{L}_p$ and $\mathcal{L}_{Adv}$ are the ℓ_p loss and adversarial loss, respectively. For simplicity, the STFT and inverse STFT (iSTFT) are omitted in the figure and the following discussion.

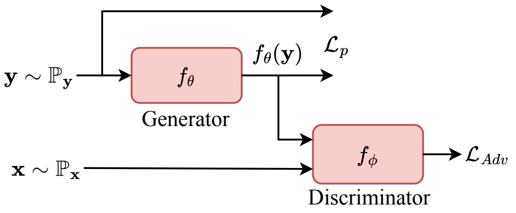


Figure 1: *Illustration of the proposed method. The noisy and clean speech, $\mathbf{y} \sim \mathbb{P}_\mathbf{y}$ and $\mathbf{x} \sim \mathbb{P}_\mathbf{x}$ are unpaired. f_θ and f_ϕ are the generator and discriminator, respectively. $\mathcal{L}_p$ and $\mathcal{L}_{Adv}$ are the ℓ_p and adversarial losses, respectively.*

The speech enhancement model is trained in an unsupervised manner only using the noisy speech based fidelity loss (i.e., $\mathcal{L}_p$ loss) and an additional adversarial loss,

$$\begin{aligned} \mathcal{L}_G &= \alpha_p \mathcal{L}_p + \mathcal{L}_{Adv}(G; D) \\ &= \alpha_p ||f_\theta(\mathbf{y}) - \mathbf{y}||_p^p - \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_\mathbf{y}} [f_\phi(f_\theta(\mathbf{y}))] \end{aligned} \tag{5}$$

where $\alpha_p > 0$ is a weight factor, $\mathcal{L}_{Adv}(G; D)$ is the adversarial loss for training the generator. It is clear that the loss in (5) is an implementation of the optimization problem in (3), and we use the Wasserstein GAN (WGAN) [17] in implementation, with which $d(\cdot, \cdot)$ corresponds to the Wasserstein distance.

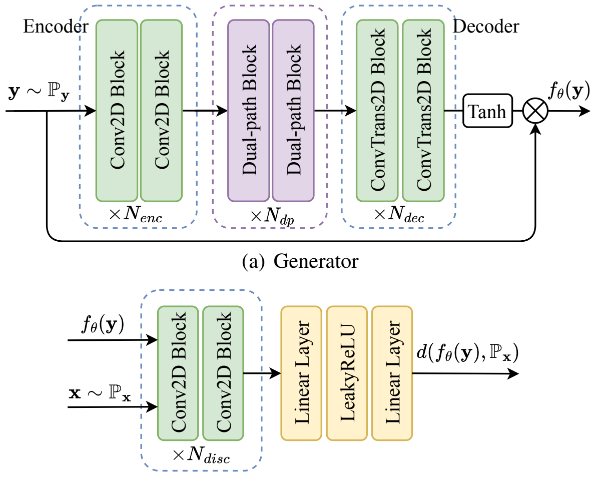
We use the WGAN loss with gradient penalty (WGAN-GP) [18] to train the discriminator,

$$\begin{aligned} \mathcal{L}_D &= \mathcal{L}_{Adv}(D; G) + \alpha_{gp} \mathcal{L}_{gp} \\ &= \mathbb{E}_{\mathbf{y} \sim \mathbb{P}_\mathbf{y}} [f_\phi(f_\theta(\mathbf{y}))] - \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_\mathbf{x}} [f_\phi(\mathbf{x})] \\ &\quad + \alpha_{gp} \mathbb{E}_{\mathbf{x}' \sim \mathbb{P}_{\mathbf{x}'}} [(||\nabla_{\mathbf{x}'} f_\phi(\mathbf{x}')|| - 1)^2], \end{aligned} \tag{6}$$

where $\mathcal{L}_{Adv}(D; G)$ is the adversarial loss for training the discriminator, $\mathcal{L}_{gp}$ is the gradient penalty, and α_{gp} is the penalty weight. $\mathbf{x}' \sim \mathbb{P}_{\mathbf{x}'}$ denotes uniform sampling along the lines between $\mathbf{y} \sim \mathbb{P}_\mathbf{y}$ and $\mathbf{x} \sim \mathbb{P}_\mathbf{x}$.

3.3. Neural Network Architecture

Figure 2 illustrates the neural network architecture of the generator and discriminator. The generator follows the U-Net architecture with N_{enc} Conv2D, N_{dp} dual-path [19], and N_{dec} ConvTrans2D blocks. The Conv2D/ConvTrans2D block includes Conv2D/ConvTrans2D and batch normalization layers followed by PReLu activation. The dual-path block [19] includes two recurrent neural network (RNN) layers that sequentially exploit the frequency feature and the time dependency. In addition, a skip connection from the input to the output is used for guiding the neural network to learn a masking other than a mapping function, and a non-linear function is adopted for bounding the mask. As for the discriminator, it contains N_{disc} Conv2D blocks and two linear layers. The Conv2D block of the discriminator has the same structure as the generator's, except spectral normalization and LeakyReLU are used. The two linear layers are also constrained by spectral normalization to make the discriminator training more stable [20]. The experimental section will introduce the hyper-parameters and training details of the neural networks.



(a) Generator

(b) Discriminator

Figure 2: *Neural network architecture. The generator uses the U-Net architecture with N_{enc} Conv2D, N_{dp} dual-path, and N_{dec} ConvTrans2D blocks. The discriminator uses N_{disc} Conv2D blocks and two linear layers.*

4. Experiments

4.1. Datasets

We conduct speech enhancement experiments on VCTK + DEMAND benchmark dataset [21]. The benchmark dataset has two training subsets, and we use the 28-speaker subset dataset that includes 11572 utterances. The test dataset includes 824 utterances from two unseen speakers. The noisy utterances in the training subset are mixed at signal-to-noise ratios (SNRs) from 0 dB to 15 dB with an interval of 5 dB, while the test subset uses different SNRs of $\{2.5, 7.5, 12.5, 17.5\}$ dB. In our experiment, we randomly select 1000 utterances from the subset as the validation dataset and use the rest, 10572, as the training dataset. The dataset is recorded at 48 kHz, and we down-sample all utterances to 16 kHz.

We conduct automatic speech recognition (ASR) experiments on the CHiME4 dataset [22]. The dataset consists of real and simulated 6-channel audio recoded/simulated in four environments (bus, cafe, pedestrian, and street). The training, validation, and test subset include 8738 utterances (7138 simulated and 1600 real), 3280 utterances (1640 simulated and 1640 real), and 2640 utterances (1320 simulated and 1320 real), respectively. In addition, for the real data of the training and validation subsets, there is a close-talking reference speech (i.e., channel 0).

4.2. Experimental Setups

4.2.1. Speech Signal

In the STFT, we adopt a hamming window with a length of 400 samples and a hop size of 100 samples to segment the speech signal, and the size of the Fourier transform is 512. The speech segment length we feed to the neural network is 2 seconds, which corresponds to 317 frames. Thus, the number of frequency bins F and frames T are 257 and 317, respectively. We concatenate the real and imaginary parts of the complex-valued STFT representation to real numbers on the channel dimension. As a result, the shape of the tensor we feed to the neural network is $B \times 2 \times F \times T$, where B is the batch size.

4.2.2. Loss Functions

Following the WGAN-GP [18], we set $\alpha_{gp} = 10$ for the discriminator loss in (6). As for the generator loss in (5), we set $p = 1$ empirically and set the hyper-parameter $\alpha_p = 10$ by grid search.

4.2.3. Neural Networks

In the generator, the Conv2D block size N_{enc} is 3, and the convolutional layers' input and output channel numbers are (2, 32), (32, 64), and (64, 128), respectively. The kernel size and stride of the three convolutional layers are set to (5, 2) and (2, 1). After applying the encoder, the shape of the tensor is $B \times 128 \times 32 \times T$. The dual-path block size N_{dp} is 2, and we use the long short-term memory (LSTM) layers to compose the block. The ConvTrans2D block size N_{dec} is 3, and the hyper-parameters of each block are set to the mirror of the encoder. In the discriminator, the Conv2D block size N_{disc} is 6, and the convolutional layers' output channel numbers are 8, 16, 32, 64, 128, and 128, respectively. The kernel size and stride of the six convolutional layers are set to (5, 2) and (2, 2), and the size of the input feature and output feature of the two linear layers are (256, 64) and (64, 1), respectively. According to the settings, the footprints of the generator and discriminator are about 1.5 million and 0.7 million, respectively.

4.2.4. Training Details

Both the generator and discriminator are initialized with Xavier and trained with Adam optimizer. The learning rate, n_{critic} in the WGAN-GP training [18], and training epoch are set to 0.0001, 10, and 1000, respectively. We implement the algorithm using the PyTorch Lightning tools [23] and train all models on a High-Performance Computing (HPC) center.

4.3. Experimental Results

4.3.1. Speech Enhancement

In the speech enhancement experiments, we use one classical MMSE-based method, four supervised learning GAN-based methods, and one unsupervised method for comparison. The classical MMSE-based method is OMLSA [24]. The four GAN-based methods are SEGAN [25], ISEGAN [26], DSEGAN [26], and SASEGAN [27], respectively. The unsupervised method is DOTN [12]. We use the wideband version PESQ (PESQ) [28], extended STOI (eSTOI) [29], scale-invariant SNR (SI-SNR), Segmental SNR (SegSNR) and DNS-MOS [30] to evaluate speech quality. In addition, composite measures for signal distortion (CSIG), noise distortion (CBAK), and overall quality (COVL), which follow the ITU-T P.835 methodology, are also used as evaluation metrics.

The audio samples of the compared methods and supplementary materials are available online[1]. Table 1 shows the evaluation results of the compared method on the VCTK + DEMAND test set, in which the highest score of each evaluation metric is shown in bold. For the classical OMLSA method, we used the Matlab source code provided by the authors to perform noise reduction on the test set. For the SEGAN and SASEGAN methods, we use the pre-trained models to estimate the denoised speech. For the DOTN method, we train the model with the source code provided by the author. For the ISEGAN and DSEGAN methods, the pre-trained models are unavailable,

[1] `https://jiang-wenbin.github.io/UnSE/`

Table 1: *Speech enhancement results of the compared methods on the VCTK + DEMAND dataset, '-' indicates that the evaluation results are not available.*

Method	Category	PESQ	eSTOI	SI-SNR (dB)	CSIG	CBAK	COVL	SegSNR (dB)	DNSMOS
Noisy	-	1.97	0.79	8.45	3.35	2.44	2.63	1.68	2.70
OMLSA [24]	Classical	2.36	0.80	**17.28**	2.62	2.87	2.40	8.52	2.72
SEGAN [25]	Supervised	2.17	0.82	16.33	3.51	2.94	2.82	7.73	2.97
ISEGAN [26]		2.24	-	-	3.23	2.95	2.69	8.17	-
DSEGAN [26]		2.35	-	-	3.55	3.10	2.93	**8.70**	-
SASEGAN [27]*		2.39	**0.84**	15.63	3.69	3.04	3.03	7.42	**3.05**
DOTN [12]	Unsupervised	2.28	0.80	14.20	2.70	2.76	2.43	6.18	2.97
Proposed	Unsupervised	**2.45**	0.80	15.96	**3.69**	**3.05**	**3.05**	7.47	2.92

* We use the SASEGAN-10 model of the last checkpoint.

and we use the evaluation results given by the authors (eSTOI, SI-SNR, and DNSMOS are not provided in the original paper).

The experimental results in Table 1 demonstrate that: *1)* the OMLSA method obtains the highest SI-SNR score; *2)* the SASEGAN method has the best performance in terms of eSTOI and DNSMOS; *3)* the proposed unsupervised method outperforms the DOTN method in most evaluation metrics and yields the highest PESQ-WB, CSIG, CBAK, and COVL scores among all comparison methods. From the results, we can conclude that the proposed unsupervised learning based speech enhancement method is superior to the DOTN and achieves competitive performance with the popular supervised learning-based method.

4.3.2. Speech Recognition

In the speech recognition experiments, we use the same neural network architecture (i.e., the generator in Figure 2) as the front-end denoising model for a fair comparison. The speech recognition model only uses the log-Mel spectrogram as an input feature, so we only operate the magnitude spectrum in the denoising model. Thus, the tanh nonlinear activation for bounding the mask in Figure 2 is replaced with the sigmoid.

We compare one supervised and three unsupervised learning methods to train the front-end denoising model. The supervised learning method used 1-channel simulated training data (7138 paired utterances) to train the denoising model (denoted by supervised (simu)). For the unsupervised learning, we trained three models: one using the same data as the supervised learning but in an unpaired manner (denoted by unsupervised (simu)), another using 6-channel real recordings of the training set as noisy speech and channel 0 recordings as unpaired clean speech (denoted by unsupervised (real)), and a third pre-trained with the previous method and fine-tuned on real recordings of the test set (denoted by unsupervised (real) + test time adaptation (TTA)). It should be noted that only the noisy speech is available for the test time adaptation, the unseen clean speech is entirely unavailable, and the model is adapted in a purely unsupervised manner.

We use the three models of Whisper [31] (i.e., tiny, base, small) for the experiments and use word error rates (WERs in percentage) as the evaluation metric. Table 2 shows the experimental results on the CHiME4 dataset. The results demonstrate that: *1)* the supervised speech enhancement method deteriorates the speech recognition performance, which is consistent with the results in [32]; *2)* as a front-end noise reduction model for speech recognition, the proposed unsupervised learning method consistently outperforms the supervised learning method; *3)* with the test time adaptation, the unsupervised method can further improve the ASR performance for the unseen test data. From the results, we can conclude that the proposed unsupervised speech enhancement method is superior to the supervised method for speech recognition, especially using the test time adaptation.

Table 2: *WERs (%) results of speech recognition on the CHiME4 dataset, TTA denotes for test time adaptation.*

Method	ASR model*	Development		Evaluation	
		simu	real	simu	real
Noisy	tiny	23.47	16.17	28.62	28.84
	base	17.07	10.44	22.60	17.92
	small	10.42	6.88	14.25	10.68
Supervised (simu)	tiny	18.76	20.87	22.24	44.53
	base	13.62	14.24	15.98	32.20
	small	10.46	13.48	12.30	29.64
Unsupervised (simu)	tiny	18.36	**17.06**	**18.61**	32.53
	base	12.37	**11.11**	**12.14**	20.71
	small	8.39	7.26	8.88	15.45
Unsupervised (real)	tiny	18.31	18.28	19.13	31.78
	base	12.96	11.63	12.87	**21.20**
	small	8.44	**7.10**	8.31	12.87
Unsupervised (real)+TTA	tiny	**18.29**	18.12	19.13	**30.76**
	base	**12.35**	11.45	13.29	21.24
	small	**8.17**	7.14	**8.25**	**11.98**

* We use the three speech recognition models (English-only) of Whisper, https://github.com/openai/whisper.

5. Conclusions

This paper proposed an unsupervised learning method for speech enhancement without requiring paired noisy-to-clean training data, which is based on the optimal transport theory. Concretely, we only use the noisy speech as the input and train the denoising model with a fidelity loss plus an adversarial loss. Meanwhile, the training objective of the discriminator is to minimize the probability distribution of the enhanced speech of the denoising model and the unpaired clean speech. The speech enhancement results on the popular benchmark dataset (VCTK + DEMAND) show that the proposed unsupervised method can achieve competitive performance with the supervised learning methods. Furthermore, the speech recognition results on the CHiME4 benchmark show that the proposed unsupervised method consistently outperforms the supervised learning method, especially using test time adaptation. Nevertheless, the proposed method is sensitive to hyper-parameters of the loss function. Future work includes studying more sophisticated models and loss functions to make the training procedure more stable.

6. References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2013.

[2] G. Hinton, L. Deng, D. Yu *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *Proc. IEEE ICASSP*. IEEE, 2019, pp. 6865–6869.

[5] Y. Hu, Y. Liu, S. Lv *et al.*, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. ISCA Interspeech*, 2020, pp. 2472–2476.

[6] W. Jiang, Z. Liu, K. Yu, and F. Wen, "Speech enhancement with neural homomorphic synthesis," in *Proc. IEEE ICASSP*. IEEE, 2022, pp. 376–380.

[7] J. Chen, Z. Wang, D. Tuo *et al.*, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *Proc. IEEE ICASSP*, 2022, pp. 7857–7861.

[8] W. Jiang, T. Liu, and K. Yu, "Efficient Speech Enhancement with Neural Homomorphic Synthesis," in *Proc. Interspeech 2022*, 2022, pp. 986–990.

[9] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. ASLP.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[10] E. Tzinis, Y. Adi, V. K. Ithapu *et al.*, "Remixit: Continual self-training of speech enhancement models via bootstrapped remixing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[11] R. E. Zezario, T. Hussain, X. Lu *et al.*, "Self-supervised denoising autoencoder with linear regression decoder for speech enhancement," in *Proc. IEEE ICASSP*. Barcelona, Spain: IEEE, May 2020, pp. 6669–6673.

[12] H.-Y. Lin, H.-H. Tseng, X. Lu, and Y. Tsao, "Unsupervised noise adaptive speech enhancement by discriminator-constrained optimal transport," in *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc., 2021, pp. 19 935–19 946.

[13] X. Lu, P. Shen, Y. Tsao, and H. Kawai, "Unsupervised neural adaptation model based on optimal transport for spoken language identification," in *Proc. IEEE ICASSP*. IEEE, 2021, pp. 7213–7217.

[14] W. Wang, F. Wen, Z. Yan, and P. Liu, "Optimal transport for unsupervised denoising learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2022.

[15] G. Peyré and M. Cuturi, "Computational optimal transport," *Foundations and Trends in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[17] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," Dec. 2017.

[18] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[19] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE ICASSP*, 2020, pp. 46–50.

[20] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=B1QRgziT-

[21] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*. ISCA, Sep. 2016, pp. 146–152.

[22] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE ASRU*. IEEE, 2015, pp. 504–511.

[23] W. Falcon and The PyTorch Lightning team, "PyTorch Lightning," 2019. [Online]. Available: https://github.com/PyTorchLightning/pytorch-lightning

[24] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Audio, Speech, Language Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[25] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. ISCA Interspeech*, 2017, pp. 3642–3646.

[26] H. Phan, I. V. McLoughlin, L. Pham *et al.*, "Improving gans for speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1700–1704, 2020.

[27] H. Phan, H. L. Nguyen, O. Y. Chen, P. Koch, N. Q. K. Duong, I. McLoughlin, and A. Mertins, "Self-attention generative adversarial network for speech enhancement," in *Proc. IEEE ICASSP*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 7103–7107.

[28] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[29] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. ASLP.*, vol. 24, no. 11, pp. 2009–2022, 2016.

[30] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE ICASSP*. IEEE, 2021, pp. 6493–6497.

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[32] S. E. Eskimez, X. Wang, M. Tang, H. Yang, Z. Zhu, Z. Chen, H. Wang, and T. Yoshioka, "Human Listening and Live Captioning: Multi-Task Training for Speech Enhancement," in *Proc. ISCA Interspeech*, 2021, pp. 2686–2690.