INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Know Your Enemy, Know Yourself: A Unified Two-Stage Framework for Speech Enhancement

Wenzhe Liu[1,2], *Andong Li*[1,2], *Yuxuan Ke*[1,2], *Chengshi Zheng*[1,2,*], *Xiaodong Li*[1,2]

[1]Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{liuwenzhe, liandong, keyuxuan, cszheng, lxd}@mail.ioa.ac.cn

Abstract

Traditional spectral subtraction-type single channel speech enhancement (SE) algorithms often need to estimate interference components including noise and/or reverberation before subtracting them while deep neural network-based SE methods often aim to realize the end-to-end target mapping. In this paper, we show that both denoising and dereverberation can be unified into a common problem by introducing a two-stage paradigm, namely for interference components estimation and speech recovery. In the first stage, we propose to explicitly extract the magnitude of interference components, which serves as the prior information. In the second stage, with the guidance of this estimated magnitude prior, we can expect to better recover the target speech. In addition, we propose a transform module to facilitate the interaction between interference components and the desired speech modalities. Meanwhile, a temporal fusion module is designed to model long-term dependencies without ignoring short-term details. We conduct the experiments on the WSJ0-SI84 corpus and the results on both denoising and dereverberation tasks show that our approach outperforms previous advanced systems and achieves state-of-the-art performance in terms of many objective metrics.

Index Terms: speech enhancement, dereverberation, unified framework, noise awareness, deep neural networks

1. Introduction

In real scenarios, clean speech tends to be contaminated by various interference components, like background noise and room reverberation [1], which may heavily impact the performance of automatic speech recognition (ASR), communication systems, and hearing aids [2]. Recent advances in deep neural networks (DNNs) have propitiated the development of speech enhancement (SE) research, where the task is formulated into a supervised problem [3, 4]. With a large scale of noisy-clean speech pairs, DNNs can gradually learn the complicated mapping from noisy feature representation to clean targets.

More recently, the performance of DNN-based SE approaches has been notably elevated with the help of advanced network topologies like CNNs [5, 6], LSTMs [7, 8], and self-attention (SA) mechanisms [9]. For these approaches, the end-to-end paradigm is usually utilized, *i.e.*, the network receives the noisy speech as the input and attempts to estimate the corresponding clean counterpart directly, as shown in Figure 1(b). Despite the satisfactory performance, they usually fall into several defects. Firstly, the network serves as the "black-box", where the denoising process implicitly lies within the network, lacking in model interpretability. Secondly, the real acoustic environments may be very complicated, *e.g.*, the noise can vary dramatically in the different acoustic scenarios and the reverberation effect may also change rapidly in the different room configurations. Therefore, it is quite difficult for the network to track these changes, which hinders the network performance.

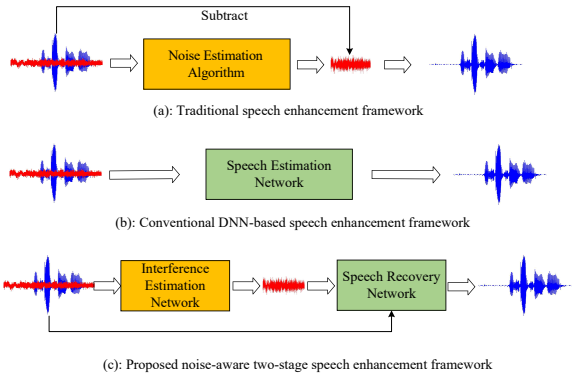

Figure 1: *Illustration of different speech enhancement frameworks. (a) Traditional spectral subtraction-type speech enhancement framework. (b) Conventional DNN-based speech enhancement framework. (c) Proposed two-stage speech enhancement framework.*

In the traditional spectral subtraction-type algorithms, noise power estimation is of paramount significance [10–13]. For these algorithms, the noise power spectral density in the time-frequency (T-F) domain is first estimated with the algorithms like MCRA [12], IMCRA [13], and U-MMSE [10]. Then the speech components can be obtained by either spectral subtraction [14] or statistical model-based gain functions [15, 16]. Spectral subtraction method is shown as an example of traditional speech enhancement algorithms in Figure 1(a).

Inspired by spectral subtraction-type SE algorithms, we propose a two-stage SE framework in this study, as shown in Figure 1(c). Different from the previous blind denoising procedure, the agnostic enhancement process is explicitly split into two pipelines, and two sub-networks are designed correspondingly, namely interference estimation network (IE-Net) and speech recovery network (SR-Net). In the first stage, we aim to estimate the interference signal coarsely with IE-Net. Note that the interference components include either background noise or room reverberation. Based on that, in the second stage, the speech estimation network aims to remove noise components and recover speech components under the guidance of interference modality. The rationale lies two-fold. Firstly, in contrast to previous DNN-based methods where the network only fo-

* corresponding author.

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-238

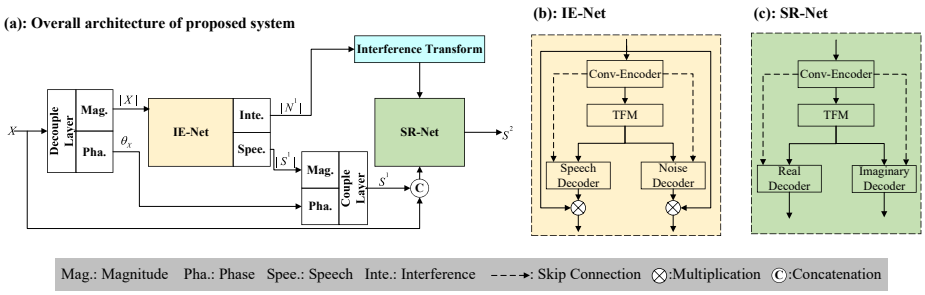

Figure 2: *Illustration of the proposed system. (a) Overall diagram of the system. (b) Detail of IE-Net. (c) Detail of SR-Net.*

cuses on speech components, the proposed network delivers the knowledge about both "**where to suppress and where to enhance**" simultaneously. As a result, the algorithm tends to be more robust toward adverse environments. Secondly, Inspired by [17–20], the curriculum learning concept is adopted [21], where the interference estimation serves as the prior information to facilitate the subsequent optimization.

To validate the effectiveness of the proposed approach, we conduct the experiments on both noise suppression and dereverberation tasks as we find that both tasks can be unified into one problem in our study. The results on the WSJ0-SI84 dataset [22] show that the proposed approach performs favorably against previous state-of-the-art systems.

The remainder of the paper is organized as follows. In Section 2, we formulate the problem. The proposed framework details are illustrated in Section 3. We present the experimental setup in Section 4. The results and analysis are given in Section 5. Some conclusions are drawn in Section 6.

2. Problem formulation

In this section, spectral subtraction and weighted prediction error (WPE) [23] algorithms are taken as the example to demonstrate the calculation of traditional SE algorithms, which can be formulated into a unified process.

Assume the noisy, the clean signals and the interference components in the T-F domain as $X_{k,l}$, $S_{k,l}$, and $D_{k,l}$, where $k \in \{1, K\}$ and $l \in \{1, L\}$ denote the index in the frequency and time axises, respectively.

When background noise serves as the interference components, spectral subtraction can be adopted. Firstly, the noise power spectral density is estimated, which is subsequently subtracted from the noisy spectrum. Take the magnitude spectral subtraction as an example, the process can be expressed as:

$$\widetilde{S}_{k,l} = \max\{(|X_{k,l}| - |\widetilde{N}_{k,l}|), 0\} \exp(j\angle X_{k,l}), \quad (1)$$

where $\widetilde{S}_{k,l}$ and $|\widetilde{N}_{k,l}|$ are the estimated speech spectral and the estimated noise magnitude, respectively.

For reverberation removal, WPE algorithm is usually deployed. Specifically, the optimal prediction filter $\mathbf{W}$ is determined to estimate the reverberation component, which is then subtracted from the original spectrum, given by:

$$\widetilde{S}_{k,l} = X_{k,l} - \mathbf{W}_k^H \bar{\mathbf{X}}_{k,l}, \quad (2)$$

where $(\cdot)^H$ stands for Hermitian transpose, $\mathbf{W}_k$ and $\bar{\mathbf{X}}_{k,l}$ are expressed as:

$$\mathbf{W}_k = [W_{k,\Delta+1}, W_{k,\Delta+2}, ..., W_{k,\Delta+P}]^T, \quad (3)$$

$$\bar{\mathbf{X}}_{k,l} = [X_{k,l-\Delta-1}, X_{k,l-\Delta-2}, ..., X_{k,l-\Delta-P}]^T, \quad (4)$$

where Δ and P are the prediction delay and the length of filter.

One can find that these two algorithms can be formulated into **a unified two-stage framework**. In the first stage, we attempt to estimate the interference components to serve as a prior reference. In the second stage, we can obtain the final target estimation with the guidance of the interference reference. Based on that, we reformulate the above procedure into a more general case, *i.e.*, the interference components are explicitly estimated from the original mixture in the first stage. Based on that, we can estimate the desired signal under this prior guidance. The whole process can be summarized as:

$$\left\{\widetilde{\mathbf{S}}^1, \widetilde{\mathbf{N}}^1\right\} = \mathcal{F}_1\left(X; \Theta_1\right), \quad (5)$$

$$\widetilde{\mathbf{S}}^2 = \mathcal{F}_2\left(X, \widetilde{\mathbf{S}}^1, \widetilde{\mathbf{N}}^1; \Theta_2\right), \quad (6)$$

where $\mathcal{F}_1\left(\cdot; \Theta_1\right)$ and $\mathcal{F}_1\left(\cdot; \Theta_2\right)$ denote the operations in the first and second stage, respectively. $(\tilde{\cdot})^1$ and $(\tilde{\cdot})^2$ refer to the estimations in the first and the second stage. Note that we estimate the interference components with the help of speech estimation, which is easier due to the fact that they are complementary tasks. Moreover, we utilize a non-linear mapping $\mathcal{F}_2\left(\cdot\right)$ to replace the subtraction operation to improve performance.

3. System description

The diagram of the proposed approach is illustrated in Figure 2(a). It consists of two pipelines and two major modules are designed accordingly, namely IE-Net and SR-Net. IE-Net is tasked with estimating interference components, and SR-Net is to remove these interference components and regenerate speech. Note that before the estimated spectrum of interference components is sent to SR-Net, it will be modified by the interference transform module (ITM) to re-encode the interference feature distribution. Specifically, IE-Net as a masking-based network obtains a better interference estimation by utilizing multi-task learning, because we believe the speech estimation is beneficial for removing the corresponding speech components of the estimated noise spectrum. As a by-product, the target magnitude estimation is then coupled with the noisy phase and then used as the input to SR-Net with noisy spectra. Finally, the interference is suppressed and an enhanced complex spectrum is generated.

3.1. IE-Net and SR-Net

The topologies of IE-Net and SR-Net are presented in Figure 2(b)-(c). One can find that both two modules have the similar structure as [24], that is, the convolutional encoder and decoder are utilized to extract the local patterns and reconstruct the spectrum, respectively. Between them, the sequence mod-

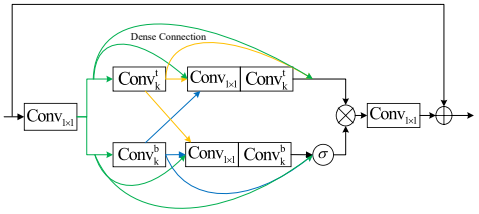

Figure 3: *Illustration of the proposed LSTDCU. Superscripts "t" and "b" denote the top and bottom branches, respectively. Subscript "k" denotes the kernel size and* 1×1 *denotes the convolution operation with* $k = 1$.

eling module is inserted to extract the long-term temporal correlations. For IE-Net, two decoders are utilized to estimate the IRM *w.r.t.* speech and interference, and the speech and interference magnitudes can be obtained by T-F filtering. For SR-Net, as we hope to estimate both real and imaginary (RI) components simultaneously, we also adopt two decoders as a result. Despite the effectiveness of temporal convolutional module (TCM) [18, 19, 25] in modeling long-range dependencies, the local-range detail tends to be blurred. To resolve this problem, we propose the temporal fusion module (TFM) which consists of a cascading of long-short term dense cross units (LSTDCU). The diagram of LSTDCU is shown in Figure 3. Compared with the previous TCM, it has several differences. (1) Instead of projecting the feature into a higher embedding space, *i.e.*, 512, we squeeze the feature size into 64, which effectively mitigates the parameter burden. (2) The dual-branch paradigm is proposed, where the information between the top and bottom branches interacts and complements with different time scales. For example, given the dilate rate in the upper branch $d^t = 2^i$, then the dilation rate in the bound version $d^b = 2^{K-i}$, where K is the number of LSTDCU in the TFM. Therefore, if the upper branch grasps the long-range time scale, then the bound branch will only model the local-range dependencies, which is beneficial for detail recovery. (3) Similar to [26], the dense connection is leveraged for both branches to grasp the information from different levels to boost the feature representation. (4) The information flows from two branches are controlled by a gating mechanism which makes this module focus on more important short-term information in long-term sequence modeling.

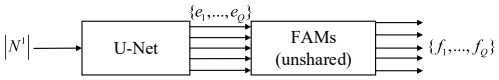

Figure 4: *Description of the proposed ITM.*

3.2. Noise Transform Module

To enable the guidance of interference estimated in the previous stage, we adopt the interference transform module (ITM) to re-transform the features, which is described in Figure 4. It includes two components, namely U-Net [27] and feature attention modules (FAMs) [28]. The U-Net is utilized here to re-transform the interference features. Then we can extract the multiple intermediate features in the decoder path, *i.e.*, $\{e_1, \cdots e_Q\} \in \mathbb{R}^{C_q \times K_q \times L_q}$, and re-weight the features in both T-F and channel dimensions by the attention mechanism. Note that C_q, K_q, and L_q denote the size of the channel, time and feature axis in the qth decoder layer output, respectively. Moreover, we introduce FAMs to pay more attention to interference-dominated bins since the interference distribution is uneven. After obtaining the weighted features $\{f_1, \cdots f_Q\}$, we feed them into the encoder parts of the SR-Net by concatenating them in the channel axis.

3.3. Loss Function

As the two-stage paradigm is utilized, we adopt the following strategy for network training. In the first stage, only IE-Net is trained, whose loss can be defined as:

$$\mathcal{L}_{IE} = \frac{1}{2}\left(\left\|S - \widetilde{M}_s X\right\|_2^2 + \left\|N - \widetilde{M}_n X\right\|_2^2\right), \quad (7)$$

where $\widetilde{M}_s$ and $\widetilde{M}_n$ denote the estimated IRMs of speech and interference, respectively.

Afterward, both IE-Net and SR-Net are jointly optimized, which can be given as:

$$\mathcal{L}_{joint} = \lambda \mathcal{L}_{IE} + \mathcal{L}_{SR}^{RI} + \mathcal{L}_{SR}^{Mag}, \quad (8)$$

$$\mathcal{L}_{SR}^{RI} = \left\|\widetilde{S}_r - S_r\right\|_2^2 + \left\|\widetilde{S}_i - S_i\right\|_2^2, \quad (9)$$

$$\mathcal{L}_{SR}^{Mag} = \left\|\sqrt{|\widetilde{S}_r|^2 + |\widetilde{S}_i|^2} - \sqrt{|S_r|^2 + |S_i|^2}\right\|_2^2, \quad (10)$$

where the superscripts "RI" and "Mag" denote the RI and Magnitude constraints. $\lambda \in [0, 1]$ controls the weight coefficient in the first stage and we set it as 0.1 empirically.

4. Experimental settings

4.1. Dataset Setup

We conduct the experiments on the WSJ0-SI84 corpus, which includes 7138 utterances by 83 speakers (42 males and 41 females). 5428, and 957 utterances by 77 speakers are chosen for training and validation, respectively. 150 utterances by another 6 untrained speakers are selected for model evaluation.

In this study, two types of interference components are explored, namely background noise and reverberation. For the first task, around 20,000 noises are randomly selected from the DNS-Challenge dataset [29] as the noise set. For each mixing process, a clean utterance and a noise vector are randomly selected and then mixed. The SNR value is randomly sampled from -5dB to 10dB. As a result, around 150,000, and 10,000 noisy-clean pairs are generated for training and validation. For testing, two challenging noises are selected, namely cafeteria and factory1 from CHiME3 [30] and NOISEX92 [31], and three SNRs are utilized, namely $\{-5\text{dB}, 0\text{dB}, 5\text{dB}\}$.

For the dereverberation task, we simulate five rooms of different sizes to generate 5040 room impulse responses (RIRs) using the image method [32]. The reverberation time T_{60} ranges from 0.3s to 1.0s. We convolve the RIR with clean utterance to obtain the reverberated speech. Totally, around 33,000, and 6,000 reverberant-direct pairs are generated for training and validation. For testing, another 1008 RIRs with a untrained room configuration is utilized with $T_{60} \in \{0.4s, 0.6s, 0.8s\}$.

4.2. Baseline Systems

We compare our proposed framework with another three advanced baselines, namely LSTM, CRN [33], and GCRN [24]. For LSTM, four LSTM layers with 1024 units are stacked, followed by a dense layer to generate the magnitude. CRN is a encoder-decoder topology with two LSTM layers serving as the bottleneck module. GCRN is the improved version of CRN, where RI components are estimated with complex spectral map-

Table 1: *Results comparison among different models for noise suppression task in terms of PESQ and ESTOI. "**Cau**" denotes whether the system is causal. **BOLD** indicates the best result in each case.*

Metrics	Cau.	PESQ				ESTOI (in %)			
SNR(dB)		-5	0	5	Avg.	-5	0	5	Avg.
Noisy	-	1.49	1.82	2.14	1.82	28.58	43.15	58.28	43.34
LSTM	✓	1.96	2.40	2.73	2.36	47.79	64.34	75.66	62.60
CRN	✓	1.99	2.44	2.79	2.41	47.87	65.10	77.06	63.34
GCRN	✓	2.12	2.63	2.97	2.58	54.55	71.33	81.00	68.96
IE-Net	✓	2.06	2.52	2.88	2.49	50.99	63.37	78.52	65.63
Proposed	✓	**2.29**	**2.81**	**3.14**	**2.75**	**60.06**	**75.41**	**83.84**	**73.10**

Table 2: *Experimental results comparisons among different speech dereverberant models in terms of PESQ and ESTOI.*

Metrics	Cau.	PESQ				ESTOI (in %)			
RT60(s)		0.4	0.6	0.8	Avg.	0.4	0.6	0.8	Avg.
Reverberant	-	2.32	2.07	1.94	2.11	59.43	46.80	38.17	48.13
LSTM	✓	2.76	2.48	2.31	2.52	76.10	68.79	63.33	69.41
CRN	✓	2.80	2.53	2.35	2.56	77.05	69.52	63.85	70.14
GCRN	✓	2.86	2.54	2.35	2.58	75.12	68.00	62.52	68.55
IE-Net	✓	2.95	2.69	2.53	2.73	79.94	73.50	69.04	74.16
Proposed	✓	**3.19**	**2.86**	**2.66**	**2.90**	**80.78**	**74.46**	**70.03**	**75.09**

ping, and all the (de)convolution layers are replaced by the gated linear units [34]. All the models are trained with the best configurations mentioned in the literatures.

4.3. Parameter Configurations

The detailed parameter configurations of our system are listed as follows. For both IE-Net and SR-Net, five convolutional blocks are utilized in the encoder, each of which includes a ConvGLU layer [24], instance normalization (IN), and PReLU. The kernel size and stride are $(2, 3)$ and $(1, 2)$ in the time and frequency axis, respectively. The decoder is the mirror version of the encoder, except all the convolution layers are replaced by deconvolutions. For TFM, 6 LSTDCUs are stacked. For the U-Net in the ITM, five convolutional blocks are utilized in the encoder and five deconvolutional blocks are taken accordingly, *i.e.*, $Q = 5$. The number of channels is 64 for convolutions.

All the utterances are sampled at 16kHz. 20ms Hanning window is utilized, with 50% overlap between adjacent frames. 320-point FFT is applied to extract the spectral features. The model is trained for 50 epochs using the Adam optimizer [35]. In the first stage, the initialized learning rate (LR) is 0.001, in the second stage, IE-Net is fine-tuned with LR = 0.0001. The batch size is set to 16 at the utterance level, where the maximum utterance length is chunked to 8s.

5. Results and analysis

In this paper, we choose perceptual evaluation speech quality (PESQ) [36] and extended short-time objective intelligibility (ESTOI) [37] as the objective metrics to compare the performance of different models.

5.1. Results comparison in the objective metrics

The noise reduction results are summarized in Table 1. Several observations can be made. Firstly, compared with magnitude-based baselines, *i.e.*, LSTM and CRN, GCRN obtains notable improvements in both metrics, indicating the significance of phase modification in the noise suppression task. Secondly, compared with previous baselines, our proposed system yields state-of-the-art performance in objective metrics. For example, from GCRN to our approach, around 0.17, and 4.14% improvements are achieved in terms of PESQ and ESTOI, indicating the superiority of our two-stage approach. Thirdly, comparing the first stage (IE-Net), we observe the two-stage system obtains considerable metric improvements, which indicates that introducing noise estimation is helpful for speech enhancement.

For the dereverberation task, one can find that our method is also significantly better than the other approaches, as shown in Table 2. From the results above, we conclude that both background noise and room reverberation can be regarded as external interference and integrated into a unified framework, where we first estimate the interference as the guidance and recover the speech complex spectrum in the second stage. Therefore, the end-to-end agnostic denoising process can be explicitly split into two steps and come with better interpretability.

Table 3: *Ablation study* w.r.t. *noise module and sequence modeling units. Proposed-1 is SR-Net without the guidance of interference.*

Systems	ID	Noise Module	Sequence Module	PESQ	ESTOI (in %)
Noisy	-	-	-	1.82	43.34
Proposed	1	✗	TFM	2.58	68.13
Proposed	2	U-MMSE	TFM	2.62	68.40
Proposed	3	IE-Net	TCM	2.71	72.48
Proposed	4	IE-Net	TFM	**2.75**	**73.10**

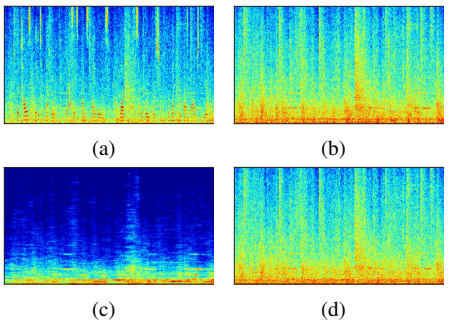

Figure 5: *Visualization of results obtained by noise estimators. (a) Noisy speech. (b) Ground truth noise. (c) Noise estimated by U-MMSE. (d) Noise estimated by IE-Net.*

5.2. Ablation analysis

We also explore the effect of interference estimation and TFM in the denoising task, as shown in Table 3. For the noise estimation module, we compare IE-Net with a traditional noise estimator named U-MMSE [10]. For the sequence learning module, we compare the proposed TFM with the TCM. Several observations can be made. Firstly, when the noise estimation module is given, the performance can be improved. Secondly, from Figure 5 we can see that the noise estimated by U-MMSE is not accurate enough which limits the performance improvement. While the proposed method can achieve better performance due to capturing more precise noise information. Thirdly, when TFM is applied, better performance is obtained than that with TCM, showing the superiority of TFM in sequence modeling.

6. Conclusions

We propose a unified framework to tackle either noise or reverberation suppression, which takes the advantages of traditional signal processing and DNNs based approaches. In all, we explicitly split the end-to-end agnostic target mapping process into two stages. In the first stage, we focus on estimating the interference components. Based on that, the desired signal can be effectively recovered under the previous guidance. Experiments on noise and reverberation suppression tasks show the superiority of our approach over previous advanced systems.

7. References

[1] C. Zheng, R. Peng, J. Li and X. Li,"A constrained mmse lp residual estimator for speech dereverberation in noisy environments," *IEEE Signal Processing Lett.*, vol. 21, no. 12, pp. 1462-1466.

[2] P. Loizou, *Speech Enhancement: Theory and Practice.* Boca Raton, FL: CRC, 2007.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 26, no. 1, pp. 1702-1726, 2018.

[4] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849-1858, 2014.

[5] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Proc. Interspeech*, pp. 3642-3646, 2017.

[6] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. MLSP*, pp. 1–6, 2017.

[7] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *J. Acoust. Soc. Amer.*, vol. 141, no. 6, pp.4705-4714, 2017.

[8] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, pp. 708–712, 2015.

[9] J. Kim, M. El-Khamy, and J. Lee, "T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement," in *Proc. ICASSP*, pp. 6649–6653, 2020.

[10] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2011.

[11] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech, Audio Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[12] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal process.*, vol. 81, no. 11, pp. 2403–2418, 2001.

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

[14] S. Kamath, P. Loizou *et al.*, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise." in *Proc. ICASSP*, pp. 164–164, 2002.

[15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, 1984.

[16] ——, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[17] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Separated Noise Suppression and Speech Restoration: LSTM-Based Speech Enhancement in Two Stages," in *Proc. WASPAA*, pp. 234–238, 2019.

[18] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "Icassp 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *Proc. ICASSP*, pp. 6628–6632, 2021.

[19] A. Li, W. Liu, C. Zheng, C. Fan and X. Li, "Two Heads are Better Than One: A Two-Stage Complex Spectral Mapping Approach for Monaural Speech Enhancement," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 29, pp. 1829–1843, 2021.

[20] R. Xu, R. Wu, Y. Ishiwaka, C. Vondrick, C. Zheng, "Listening to Sounds of Silence for Speech Denoising," *Proc. NeurIPS*, pp. 9633–9648, 2020.

[21] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. ICML*, pp. 41–48, 2009.

[22] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *DARPA Speech and Language Workshop*. Morgan Kaufmann Publishers, 1992.

[23] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, 2010.

[24] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 28, pp. 380–390, 2020.

[25] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[26] G. Huang, S. Liu, L. Van der Maaten, and K. Q. Weinberger, "Condensenet: An efficient densenet using learned group convolutions," in *Proc. CVPR*, pp. 2752–2761, 2018.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, pp. 234–241, 2015.

[28] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "Ffa-net: Feature fusion attention network for single image dehazing," arXiv preprint arXiv:1911.07559, 2019.

[29] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," arXiv preprint arXiv:2009.06122, 2020.

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

[31] A. Varga and J. M. H. Steeneken, "Assessment for automatic speech recognition ii: Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247251, 1993.

[32] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[33] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Proc. Interspeech*, pp. 3229–3233, 2018.

[34] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. ICML*, pp. 933–941, 2017.

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

[36] 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 *Proc. ICASSP*, pp.749-752, 2001.

[37] J. Jensen, and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio. Speech, Lang. Process.*, vol. 24, no. 11, pp.2009-2022, 2016.