A TIME DOMAIN PROGRESSIVE LEARNING APPROACH WITH SNR CONSTRICTION FOR SINGLE-CHANNEL SPEECH ENHANCEMENT AND RECOGNITION

Zhaoxu Nian[1], *Jun Du*[1,*], *Yu Ting Yeung*[2], *Renyu Wang*[2]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]Huawei Noah's Ark Lab, China
zxnian@mail.ustc.edu.cn, ✉jundu@ustc.edu.cn, yeung.yu.ting@huawei.com, wangrenyu1@huawei.com

ABSTRACT

Single-channel speech enhancement for automatic speech recognition (ASR) has been widely studied. However, most speech enhancement methods conduct over suppression and introduce distortion, which limits performance gains or even deteriorates the back-end performance. The key to solving this problem is preserving the integrity of speech while suppressing the background noises. Therefore, we propose a time domain progressive learning (TDPL) approach for speech enhancement and ASR. TDPL model consists of encoder, progressive enhancer and decoder. Both SNR-increased intermediate target with less speech distortion and clean target with better listening quality/intelligibility are learned, which are provided for ASR pre-processing and speech communication, respectively. Additionally, we also present an SNR constriction loss that is fit for TDPL to further improve ASR performance. We evaluate the proposed methods on CHiME-4 real evaluation set. The results show that the TDPL method significantly outperforms time domain speech enhancement methods and frequency domain progressive learning methods in ASR task, and the intermediate output of TDPL achieves a 36.3% relative word error rate reduction with a powerful ASR back-end without retraining. Moreover, the estimated clean output achieves certain improvement on CHiME-4 simulation evaluation set in terms of PESQ and STOI measures.

Index Terms— Time domain, progressive learning, SNR constriction, automatic speech recognition, speech enhancement

1. INTRODUCTION

Automatic speech recognition (ASR) has achieved great improvement with the development of deep learning methods [1, 2]. Nonetheless, even for an ASR system trained on multi-conditional training data, background noise and reverberation in realistic conditions can severely deteriorate its performance.

Speech enhancement [3] aims to attenuate background noise in a given noisy speech. It is widely used for speech communication. Moreover, it can be used as a front-end system to improve performance and robustness of ASR systems [4, 5]. In the last few years, deep neural networks (DNN) based speech enhancement has been extensively studied. Most of them enhance speech in frequency domain [6, 7] by applying short-time Fourier transform (STFT) [8]. However, speech enhancement in frequency domain has inevitable shortcomings. First, clean phase information is often discarded which is necessary for speech recognition. Second, some training targets such as ideal ratio mask (IRM) [6] cannot reconstruct speech signals perfectly. Finally, computational cost of time-frequency transform limits the practicability of frequency domain methods [9].

Recently, time domain front-end approaches including speech enhancement [10, 11, 12] and separation [13, 14] attract more research attentions. However, most time domain speech enhancement methods focus on enhancement measures for speech communication such as perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) and only few studies discuss time domain speech enhancement for ASR. In [15], convolutional time-domain audio separation network (Conv-TasNet) [14] is introduced to speech enhancement for ASR and improves ASR performance on CHiME-4 data [16]. However, [15] trained the model on extra simulated data with reverberation instead of the official training set. In fact, speech enhancement is difficult to directly improve ASR performance. The reason is that speech enhancement methods usually take IRM, clean log-power spectra [7] or clean waveform as the training target, which will cause over suppression that is harmful to ASR performance [17]. Thus it is quite challenging for front-end technologies to yield performance gains on acoustic models using multi-condition training without retraining [18]. Some researchers have proposed methods to alleviate the problem of over suppression. In [17], asymmetric loss is proposed to improve speech preservation ability. In [19], the estimated IRM is incorporated into the procedure of improved minima controlled recursive averaging (IMCRA) [20] approach to avoid data mismatch between the training set and test set. In [21], SNR-based progressive learning for speech enhancement is proposed, which divides a whole network into stacking blocks and forces them to gradually learn less-noisy spectral features in a progressive manner until it reaches the clean spectral features. Experiments show that the intermediate target of progressive learning always improves ASR performance because it conducts a trade-off between speech distortion and preservation [22].

In this study, we propose a novel time domain progressive learning (TDPL) approach for speech enhancement and recognition. Our main contributions are in three aspects: First, we propose an SNR-based progressive learning network in time domain. Second, the proposed model achieves satisfactory performance in metrics for both speech communication and ASR, which improves its potential for different applications. The proposed method can significantly improve ASR performance of a strong multi-condition training back-end without retraining acoustic model. Finally, an SNR constriction loss is proposed to utilize the SNR relationship between different targets. We conduct the experiments on CHiME-4 test sets. The results show that the TDPL method outperforms the frequency domain progressive learning (FDPL) method and time domain speech enhancement method. TDPL achieves a 36.3% relative word error rate (WER) reduction over the system with unprocessed noisy speech.

*corresponding author

2. REVIEW OF PROGRESSIVE LEARNING

In this section, the key principle of progressive learning in frequency domain [21] is briefly introduced. Traditional deep learning based speech enhancement methods often take noisy spectral features as input and clean spectral features or masks as targets. However, it is hard to learn this complex relationship accurately for neural networks, and no one knows what the networks learn in the middle layers. In the progressive learning method, the whole training procedure is decomposed into multiple sub-training stages learning different spectra of gradually increasing SNRs as shown in Fig. 1.

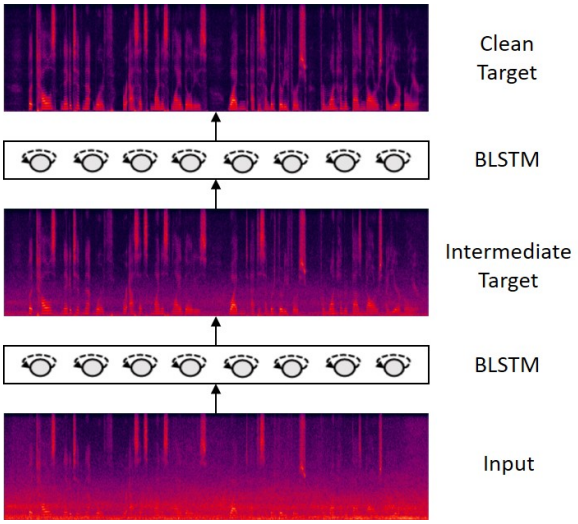


Fig. 1. An illustration of frequency domain progressive learning

The input of the whole network is the log-power spectra (LPS) of low-SNR speech. Each bi-directional long short-term memory (BLSTM) layers take the output of the preceding layer as input and the sub-training target is the LPS of the speech with a higher SNR than the previous layers. Through this method, we can clearly know that the SNRs of the generated speech increase through each layer of the network until we finally obtain a clean speech. The final clean output of progressive learning method can achieve good results in speech enhancement metrics (e.g., PESQ and STOI) because the decomposition of the whole problem makes the learning of each submodule easier. Compared with the clean output, the intermediate output can achieve better results in the metric of speech recognition. Robust ASR systems are usually trained with multi-conditional noisy training data which makes the systems have ability to deal with noises. Thus clean output is not necessary for ASR, and in most cases, it will even degrade the performance of noise-robust ASR. Intermediate output conducts a trade-off between speech distortion and speech preservation, which leads to satisfactory ASR performance improvement.

3. TIME DOMAIN PROGRESSIVE LEARNING

The system flowchart of the proposed TDPL is shown in Fig. 2 (A). The whole network consists of three parts: encoder, progressive enhancer and decoder. A learned encoder is used to replace STFT to extract high-dimensional speech features with phase information. One enhancer takes the features as input and generates mask under the guidance of the speech with higher SNR than input speech. Then, the other enhancer takes the output features of the previous enhancer as input and estimates the clean mask. Finally, masked encoder features are used to reconstruct two speech signals through two learned decoders. The intermediate output is used for ASR systems and the clean output is used for speech communication.

3.1. Problem description

Speech enhancement method aims to suppress noise from noisy speech. Consider $x(t)$ is a noisy speech corrupted by additive background noises, which can be defined as follows:

$$x(t) = s_c(t) + d(t) \tag{1}$$

where $s_c(t)$ is the clean target speech, $d(t)$ is the background noise. However, enhanced speech is not required to be completely clean when we use speech enhancement as the front-end of ASR system. The intermediate target for ASR system is defined as

$$x(t) = s_n(t) + d'(t) \tag{2}$$

where $s_n(t)$ is the intermediate target speech while $d'(t)$ is the noise that can be reduced by enhancement system. The remaining noise in $s_n(t)$ can be processed by multi-condition training ASR back-end. We aim to recover the clean target $s_c(t)$ and the intermediate target $s_n(t)$ in one model.

3.2. Time domain progressive learning network

TDPL network mainly consists of three parts: encoder, progressive enhancer and decoder. Given the samples vector $\mathbf{x}$ of a noisy speech, the learned encoder takes place of STFT to transform the time domain speech to high dimensional representation:

$$\mathbf{X} = \text{encoder}(\mathbf{x}) \tag{3}$$

where encoder$(\cdot)$ is the encoder function and $\mathbf{X}$ is the noisy speech features. Learned encoder can utilize adequate speech information including phase information.

The structure of progressive enhancer shown in Fig. 2(B) is similar to the separator in Conv-TasNet [14]. We design two enhancers to progressively increase SNR of input speech and estimate two masks of different targets. Each enhancer mainly consists of two stacks of n 1-D Conv blocks with dilation factor 1, 2, 4, ..., 2^{n-1}, which is the same as the configuration in Conv-TasNet. For the first enhancer designed for intermediate target, the cleaner features with a specific SNR gain is generated after 1-D Conv stacks and intermediate mask is estimated:

$$[\mathbf{X}_i, \mathbf{M}_n] = \text{enhancer}(\mathbf{X}) \tag{4}$$

where $\mathbf{X}_i$ represents the speech features input to the next enhancer, $\mathbf{M}_n$ is the mask of the intermediate target. enhancer$(\cdot)$ represents the processing of enhancer. For the next enhancer, the cleaner features $\mathbf{X}_i$ are adopted as the input to estimate the clean mask:

$$[\_, \mathbf{M}_c] = \text{enhancer}(\mathbf{X}_i) \tag{5}$$

where _ means that the last enhancer does not output features. $\mathbf{M}_c$ represents the clean mask.

Finally, decoders reconstruct the processed speech from the estimated masks and the noisy input features using 1-D transposed convolution operation:

$$\hat{\mathbf{s}}_n = \text{decoder}(\mathbf{X} \odot \mathbf{M}_n) \tag{6}$$

$$\hat{\mathbf{s}}_c = \text{decoder}(\mathbf{X} \odot \mathbf{M}_c) \tag{7}$$

where decoder$(\cdot)$ represents the decoder function and $\odot$ denotes element-wise multiplication. $\hat{\mathbf{s}}_n$ and $\hat{\mathbf{s}}_c$ represent the estimated intermediate target and clean target, respectively.

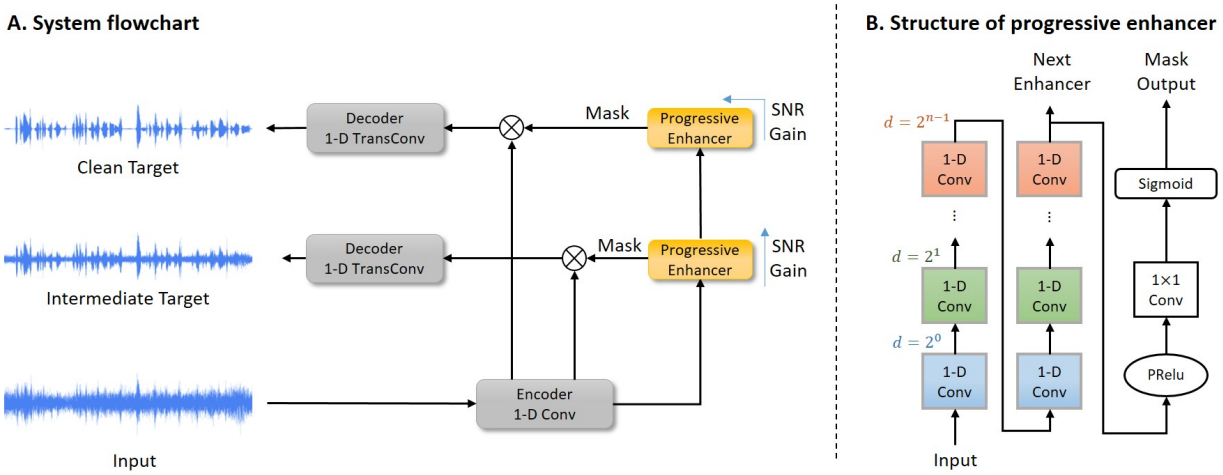


Fig. 2. (A): The flowchart of proposed time domain progressive learning method. (B): The design of progressive enhancer block

3.3. Loss functions

We first employ classic SNR loss to train TDPL model. SNR loss preserves the scale of target signal and prevents the performance of ASR from decreasing due to amplitude change [15]. In TDPL model, the clean output and intermediate output calculate the SNR with the corresponding target. The progressive SNR loss can be defined as:

$$E = -\eta_c \text{SNR}(\mathbf{s}_c, \hat{\mathbf{s}}_c) - \eta_n \text{SNR}(\mathbf{s}_n, \hat{\mathbf{s}}_n) \quad (8)$$

where η_c and η_n are the weighting factors of clean target loss and intermediate target loss, respectively. SNR loss is used instead of scale-invariant signal-to-noise ratio (Si-SNR) [23] loss because Si-SNR ignores the influence of estimated speech amplitude which might be harmful to ASR system. Moreover, preliminary experiments show that using SNR loss instead of Si-SNR, the intermediate output can achieve better ASR performance.

Besides, SNR loss with SNR constriction is proposed to fully utilize the SNR relationship among the input noisy speech, the intermediate target and the clean target. Assuming background noise is only additive noise, according to Eqs. (1) and (2), the relationship between different speech signals and noise is shown in Fig.3.

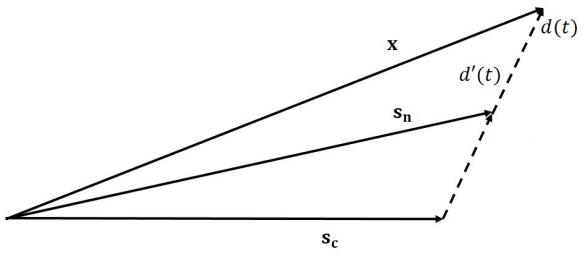


Fig. 3. An illustration of SNR relationship between targets and input

In the training set, the intermediate target and input noisy speech are simulated by clean target corrupted by the same noise of different amplitude. To remain this relationship in training procedure, a constrict item is added to SNR loss. The SNR constriction loss is defined as:

$$E = \text{SNR}loss + \lambda||\text{norm}(\hat{\mathbf{s}}_n - \mathbf{s}_c) - \text{norm}(\mathbf{x} - \mathbf{s}_c)||_2^2 \quad (9)$$

where SNR$loss$ denotes the SNR loss described in Eq.(8). λ is the weighting factor. norm$(\cdot)$ represents the normalization operation. The estimated $\hat{\mathbf{s}}_n$ is restricted to be on the dotted line of Fig.3. Through this loss, the SNR information of intermediate target can be preserved to the greatest extent and the estimated speech is prevented from producing unknown distortion.

4. EXPERIMENTS

4.1. Data corpus

Clean speech derived from WSJ0 corpus is corrupted with CHiME-4 noise [16] at three SNR levels (-5dB, 0dB, 5dB) to build a 36-hour training set as input noisy speech. The intermediate target is built by the same clean speech and noise with 10dB SNR gain (5dB, 10dB, 15dB) compared with noisy speech. We present the experimental evaluation of TDPL on the CHiME-4 official evaluation set which includes 1320 real and 1320 simulated recordings in four conditions: bus (BUS), cafe (CAF), pedestrian area (PED), and street (STR).

4.2. Implementation details

For TDPL model, each stack of 1-D Conv blocks in progressive enhancer has 6 1-D Conv layers and the remaining hyper-parameters setting is similar to original Conv-TasNet [14]. Besides, we do not use skip connection in progressive enhancer. We employed PyTorch to train the TDPL network. The model was trained for 30 epochs. The learning rate was initialized as 0.001 using Adam optimizer [24]. The batch size was 4. The loss weight parameter η_c and η_n were 1. The weight factor λ in SNR constriction loss was 2.

We trained three additional models to compare with proposed method. The first model is an FDPL model that follows the progressive learning neural network configuration in [22]. The second one is a conventional speech enhancement Conv-TasNet denoted as TDSE and directly learns clean waveform using SiSNR loss. The model structure of TDSE is almost same as TDPL model. In order to let the network directly learn the mapping to clean target, we discard the middle decoder and target. The last model is TDSE trained with noise reconstruction (NR) loss proposed in [15].

For ASR system, a TDNN-based back-end is adopted to evaluate TDPL approach without acoustic models retraining. The acoustic

model is TDNN with LF-MMI training. The language models are 5-gram with Kneser-Ney (KN) smoothing for the first-pass decoding [25] and the simple RNN-based language model for rescoring.

4.3. Results and analysis

4.3.1. Effect of TDPL on ASR performance

Table 1. WER (%) comparison TDPL on CHiME-4 real test set.

Target	Loss	BUS	CAF	PED	STR
unprocessed	unprocessed	21.26	13.37	10.37	8.87
clean	SNR	17.70	14.03	13.10	8.05
+10dB	SNR	12.92	**8.52**	7.81	5.98
+10dB	SNR constriction	**12.85**	8.59	**7.20**	**5.62**

Table 1 shows the WER results of proposed TDPL method on the four environments of CHiME-4 real test set. "unprocessed" refers to the case where the real noisy speech is directly fed into the recognition system. "+10dB" represents the intermediate output achieves 10dB SNR gain compared to the input noisy speech.

By comparing the WER of clean output with the noisy speech, clean output can achieve obvious improvement in the most adverse environment. However, in several less-noisy environments, clean output yields limited improvement or even degrades ASR performance. "+10dB" intermediate output trained with SNR loss performs significant improvement in all environments compared to unprocessed speech and clean output, which indicates over suppression caused by learning clean waveform or spectral features is the main problem for speech enhancement systems to ASR performance improvements. The design of intermediate target alleviates the over suppression problem and leads to improve ASR performance. Moreover, SNR constriction loss outperforms SNR loss mainly in less-noisy environments "PED" and "STR", which demonstrates that remaining the SNR relationship decreases the distortion in less-noisy environments.

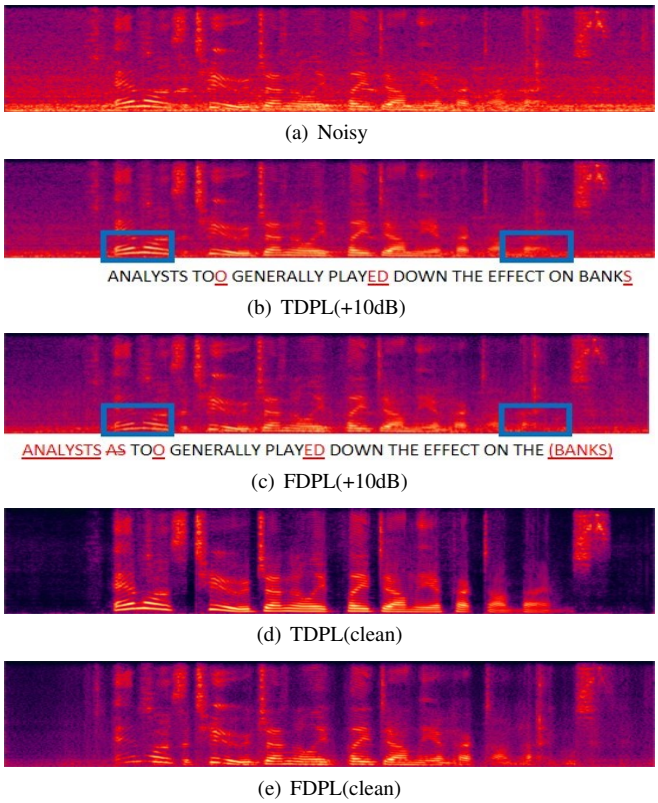
4.3.2. Comparison of TDPL and other methods

Table 2. Average WER, PESQ and STOI comparison of different approaches on CHiME-4 real or simulation test set.

System	Real data	Simu data	
	WER	PESQ	STOI
Noisy	13.44	1.976	0.811
FDPL	12.48	2.344	0.845
TDSE	12.82	2.657	0.904
TDSE(NRloss) [15]	11.99	2.599	0.903
TDPL	**8.56**	**2.668**	**0.908**

Table 2 shows the average WER, PESQ and STOI results of TDPL and other methods on CHiME-4 test set. It should be noted that real test set is used to evaluate WER while the simu test set is adopted to evaluate enhancement metrics PESQ and STOI. In FDPL and TDPL, the intermediate output and clean output are used for the corresponding tasks. TDPL outperforms all the other methods and achieves a relative WER reduction of 36.3% compared to unprocessed noisy speech. TDPL significantly improves both the enhancement and ASR metrics while the model size and computational cost are similar to TDSE. This phenomenon shows the practicability of TDPL method in realistic application.

In Fig. 4, a representative sample utterance recorded in real adverse BUS environment is selected to intuitively compare TDPL and FDPL. In the blue box region of Fig. 4 (c), the target speech is more suppressed and distorted, leading to substitution errors in the corresponding ASR results. The intermediate output of TDPL in Fig. 4 (b) achieves stronger speech preservation ability. Compared the clean output of TDPL and FDPL, it is obvious that TDPL strongly suppresses noise which leads to better enhancement performance.



(a) Noisy

(b) TDPL(+10dB)

(c) FDPL(+10dB)

(d) TDPL(clean)

(e) FDPL(clean)

Fig. 4. An example of spectrogram and ASR results of time domain and frequency domain PL. (a) Noisy speech, (b) +10dB intermediate output of TDPL, (c) +10dB intermediate output of FDPL, (d) clean output of TDPL, (e) clean output of FDPL

5. CONCLUSION

In this paper, we propose a TDPL approach for speech enhancement and recognition in one lightweight model. TDPL model consists of encoder, progressive enhancer and decoder, which progressively learns two targets of different SNRs. Over suppression problem is alleviated by introducing the intermediate target. Compared with FDPL and time domain enhancement methods, TDPL method has achieved promising results in both ASR and enhancement metrics. Besides, an SNR constriction loss is proposed to further improve ASR performance. The intermediate output of TDPL achieves over 30% WER reduction when compared to unprocessed noisy speech using a powerful ASR back-end without retraining. The positive experimental results demonstrate the effectiveness of TDPL approach.

6. ACKNOWLEDGE

This work was supported by the National Natural Science Foundation of China under Grant No. 62171427.

7. REFERENCES

[1] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4845–4849.

[2] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 conversational speech recognition system," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5934–5938.

[3] Vladimir Botchev, "Speech enhancement: Theory and practice (2nd ed.)," *Computing reviews*, vol. 54, no. 10, pp. 604–605, 2013.

[4] Yan Hui Tu, Jun Du, Lei Sun, Feng Ma, and Chin Hui Lee, "On design of robust deep models for CHiME-4 multi-channel speech recognition with multiple configurations of array microphones," in *Interspeech 2017*, 2017.

[5] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5024–5028.

[6] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[7] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[8] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[9] Ashutosh Pandey and DeLiang Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[10] Dario Rethage, Jordi Pons, and Xavier Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5069–5073.

[11] Szu-Wei Fu, Tao-Wei Wang, Yu Tsao, Xugang Lu, and Hisashi Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[12] Ashutosh Pandey and DeLiang Wang, "Dense CNN with self-attention for time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1270–1279, 2021.

[13] Yi Luo and Nima Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 696–700.

[14] Yi Luo and Nima 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.

[15] Keisuke Kinoshita, Tsubasa Ochiai, Marc Delcroix, and Tomohiro Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7009–7013.

[16] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[17] Alan Chiao, Alex Gruenstein, Ignacio Lopez Moreno, Jason Pelecanos, Kevin William Wilson, Marily Nika, Mert Saglam, Quan Wang, Renjie Liu, Wei Li, and Yanzhang (Ryan) He, "Voicefilter-lite: Streaming targeted voice separation for on-device speech recognition," in *Interspeech*, 2020.

[18] Hao Tang, Wei-Ning Hsu, Francois Grondin, and James Glass, "A study of enhancement, augmentation, and autoencoder methods for domain adaptation in distant speech recognition," 2018.

[19] Y. Tu, J. Du, and C. Lee, "Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2080–2091, 2019.

[20] I Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[21] Y. Tu, J. Du, T. Gao, and C. Lee, "A multi-target snr-progressive learning approach to regression based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1608–1619, 2020.

[22] Zhaoxu Nian, Yan-Hui Tu, Jun Du, and Chin-Hui Lee, "A progressive learning approach to adaptive noise and speech estimation for speech enhancement and noisy speech recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6913–6917.

[23] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey, "SDR – half-baked or well done?," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[24] Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization," *arXiv e-prints*, p. arXiv:1412.6980, Dec. 2014.

[25] R. Kneser and H. Ney, "Improved backing-off for M-gram language modeling," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, 1995, vol. 1, pp. 181–184 vol.1.