ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413395

A PROGRESSIVE LEARNING APPROACH TO ADAPTIVE NOISE AND SPEECH ESTIMATION FOR SPEECH ENHANCEMENT AND NOISY SPEECH RECOGNITION

Zhaoxu Nian[1], *Yan-Hui Tu*[1], *Jun Du*[1*], *Chin-Hui Lee*[2]

[1]University of Science and technology of China, Hefei, Anhui, P. R. China
[2]Georgia Institute of Technology, Atlanta, GA. USA
zxnian@mail.ustc.edu.cn, yanhuitu@ustc.edu.cn, ✉jundu@ustc.edu.cn, chl@ece.gatech.edu

ABSTRACT

In this paper, we propose a progressive learning-based adaptive noise and speech estimation (PL-ANSE) method for speech preprocessing in noisy speech recognition, leveraging upon a frame-level noise tracking capability of improved minima controlled recursive averaging (IMCRA) and an utterance-level deep progressive learning of nonlinear interactions between speech and noise. First, a bi-directional long short-term memory model is adopted at each network layer to learn progressive ratio masks (PRMs) as targets with progressively increasing signal-to-noise ratios. Then, the estimated PRMs at the utterance level are combined within a conventional speech enhancement algorithm at the frame level for speech enhancement. Finally, the enhanced speech based on multi-level information fusion is directly fed into a speech recognition system to improve the recognition performance. Experiments show that our proposed approach can achieve a relative word error rate (WER) reduction of 22.1% when compared to results attained with unprocessed noisy speech (from 23.84% to 18.57%) on the CHiME-4 single-channel real test data.

Index Terms— Speech recognition, speech enhancement, progressive learning, improved minima controlled recursive averaging, adaptive noise and speech estimation

1. INTRODUCTION

Automatic speech recognition (ASR) [1] is developing rapidly and widely used in our daily lives [2]. Nonetheless, in realistic adverse environment, the speech signal is corrupted by complicated noises and interferences, which may degrade the performance of ASR systems. Speech enhancement [3] aims to suppress the background noise from the observed noisy speech. It is usually used as a front-end system to improve the performance and robustness of ASR systems [4].

Supervised single-channel speech enhancement based on deep neural network (DNN) [5, 6, 7] have been proposed to improve the performance of ASR [8, 9, 10] by many researchers. However, the acoustic models in an ASR system usually need to be retrained to take advantage of speech enhancement [11] because there exists mismatch between enhanced speech and the data used in training acoustic models [12]. Moreover, directly learning from ideal ratio mask (IRM) [7] or clean log-power spectra (LPS) may distort the speech spectra. This is often harmful to ASR system accuracies. Therefore, it is quite challenging for pre-processing approaches to yield performance gains on acoustic models using multi-condition training without retraining [13]. In [6, 14], a long short-term memory recurrent neural network (LSTM-RNN) was proposed for speech enhancement as opposed to conventional DNN-based speech enhancement, and achieved good noise reduction at low signal-to-noise ratios (SNRs). In [15, 16], a SNR-based progressive learning approach was proposed to decompose the whole training process into multiple sub-training stages, each corresponding to different SNRs targets. In [17], a progressive multi-target model was proposed as a pre-processing neural network to improve ASR performance.

The classical approaches use statistical signal processing to enhance speech [18, 19]. They have abilities to adapt to the noise level and perform well with quasi-stationary noises but non-stationary noises can not be tracked accurately. Deep learning-based speech enhancement algorithms typically perform better than a classic noise suppressor in some normal cases. However, it may not work well when there exists mismatch between training and test data. Classical speech enhancement algorithms and deep learning-based single-channel speech enhancement approaches have demonstrated different strengths and weaknesses. Some researchers propose to combine the advantages of the classic noise suppression and DNN-based mask estimation methods. In [20], IRM and clean LPS estimated by LSTM were mixed with traditional suppression rule to calculate speech spectrum. In [21], a learning target, namely improved speech presence probability (ISPP), was proposed by incorporating the estimated IRMs from a teacher model into the procedure of improved minima controlled recursive averaging (IMCRA) [22] approach and better ASR performances were obtained.

In this study, based on our previous work [21], we propose a novel progressive learning-based adaptive noise and speech estimation (PL-ANSE) framework to further combine the advantages of the estimated progressive ratio masks (PRMs) from progressive learning-based framework and conventional IMCRA to achieve satisfactory ASR results without retraining acoustic models. First, a bi-directional long short-term memory (BLSTM) progressive learning model is introduced to estimate PRMs at an utterance level. Then, the estimated PRMs are incorporated into the IMCRA procedure. A new gain function combined with PRMs and the IMCRA gain function is proposed to recover speech frame by frame. Moreover, our proposed approach has the ability to adaptively control the tradeoff between noise reduction and speech distortion and estimate noise accurately by utilizing the information provided by the PRMs. Experiments on the CHiME-4 real test sets show that the PRMs estimated by the first intermediate layer of the PL model achieve good ASR performance when compared to that obtained with unprocessed noisy speech without acoustic model retraining, with a relative word error rate (WER) reduction of 14.0% averaging over the real test set. While our proposed approach outperforms the PRMs with an additional WER reduction of 9.4% or an over relative WER reduction of 22.1% from WERs obtained with noisy speech.

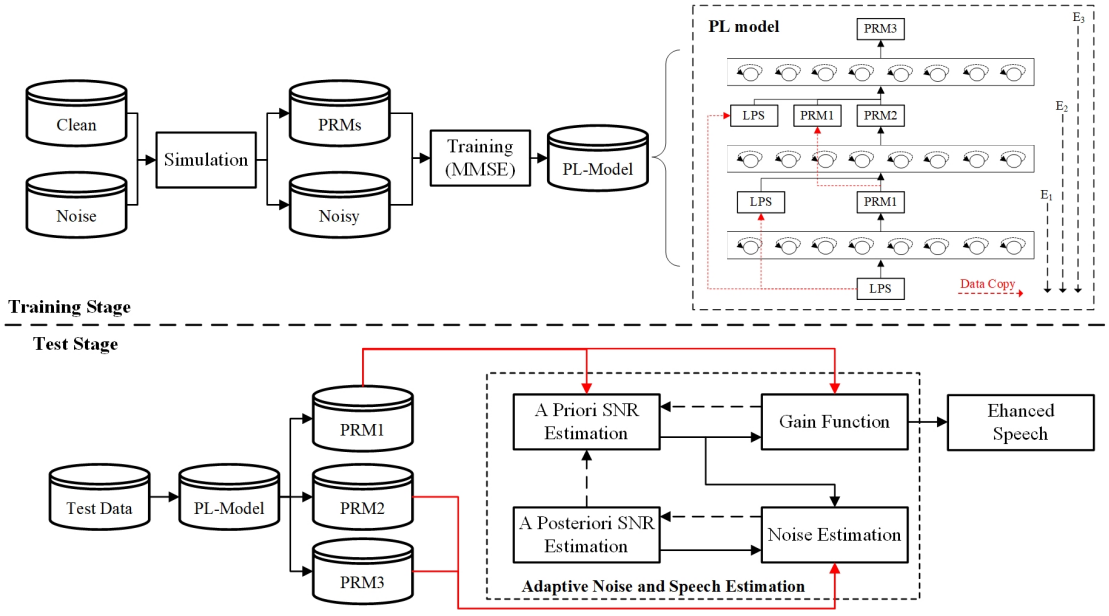


Fig. 1. The framework of the proposed approach.

2. SPEECH ENHANCEMENT PREPROCESSING

The sketch map of our proposed speech enhancement method is shown in Fig. 1. In training stage, BLSTM is adopted at each layer to learn PRMs using MMSE criterion as shown in the upper part of Fig. 1. In test stage, estimated masks are merged with the updating procedure of IMCRA, corresponding to the red line in Fig. 1. The part in the dotted box updates by frame with utterance-level estimated masks to achieve multi-level speech enhancement.

2.1. Utterance-level progressive learning

In [17], the progressive ratio mask (PRM) is adopted in the progressive learning framework based on LSTM, which demonstrates the effectiveness for speech recognition. In this paper, the BLSTM, which outperforms the LSTM at a whole utterance level, is adopted as the regression model to construct the progressive learning framework, denoted as BLSTM-PL. But it should be noted that BLSTM is not the only choice. For instance, we can use LSTM to make the system causal. The log-power spectra (LPS) feature is adopted as the input and the output is clean ideal ratio mask (IRM). A series of PRMs are the learning targets of intermediate layers. PRM, to perform a tradeoff between noise reduction and speech distortion, is defined as:

$$M_{\mathrm{PRM}_m}(k,l) = \frac{|S(k,l)|^2 + |N_{\mathrm{m}}(k,l)|^2}{|S(k,l)|^2 + |N(k,l)|^2} \tag{1}$$

where $S(k,l)$ represents the short-time Fourier transform (STFT) of the speech signal at the time frame l and frequency bin k. $N_{\mathrm{m}}(k,l)$ and $N(k,l)$ represent the STFT of the noise in one PRM target and input signals at the T-F unit (k,l) respectively. When $N_{\mathrm{m}}(k,l) = 0$, $M_{\mathrm{PRM}_m}(k,l)$ is equal to the traditional IRM $M_{\mathrm{IRM}}(k,l)$. When optimizing the parameters in Fig. 1, we present a weighted MMSE criterion with M target layers to calculate loss.

$$E = \sum_{m=1}^{M} \eta_m E_m \tag{2}$$

$$E_m = \frac{1}{N}\sum_{k,l} ||\mathcal{F}_m(\hat{M}_{\mathrm{PRM}_1}(k,l), \hat{M}_{\mathrm{PRM}_2}(k,l), ..., \hat{M}_{\mathrm{PRM}_{m-1}}(k,l), \Lambda_m) - M_{\mathrm{PRM}_m}(k,l)||_2^2 \tag{3}$$

where η_m is the weighting factor for m^{th} target layer. Λ_m is the set of weight matrices and bias vectors. $M_{\mathrm{PRM}_m}(k,l)$ denotes the target mask of m^{th} layer. $\mathcal{F}_m(\theta)$ is the neural network output of m^{th} target layer. Here M is 3 and $M_{\mathrm{PRM}_3}(k,l)$ is equal to $M_{\mathrm{IRM}}(k,l)$.

To train the BLSTM-PL model, large amounts of training data simulated by different noise types and SNRs are used to improve the generalization ability of the model. However, if there exists a high mismatch between training data and test data, the BLSTM-PL model can not perform well. Moreover, the BLSTM-PL model can not adjust its parameters frame by frame for test data. While for speech recognition, the key to front-end speech enhancement is to greatly reduce the noise in the non-speech segments and keep the speech information as much as possible in the speech segments.

2.2. Frame-level IMCRA

In this section, the key principle of the IMCRA approach [22] is briefly introduced. For the IMCRA method, the key point is the estimation of *a priori* SNR $\xi(k,l)$ and *a posteriori* SNR $\gamma(k,l)$, which can be defined as:

$$\xi(k,l) \triangleq \frac{\lambda_s(k,l)}{\lambda_d(k,l)} \tag{4}$$

$$\gamma(k,l) \triangleq \frac{|X(k,l)|^2}{\lambda_d(k,l)} \tag{5}$$

where $\lambda_s(k,l)$ and $\lambda_d(k,l)$ denote the spectrum of target speech and noise, respectively. $X(k,l)$ denotes the STFT of noisy speech signal. For estimating *a posterior* SNR, only the noise is necessary to be estimated, by initializing $\lambda_d(k,l)$ at the first frame with $\lambda_d(k,l) = |X(k,0)|^2$. Then, $\lambda_d(k,l+1)$ is calculated by a recur-

sive averaging between $\lambda_d(k,l)$ and $|X(k,l)|^2$ as follow:

$$\begin{aligned}\lambda_d(k,l+1) &= \lambda_d(k,l)p(k,l) \\ &+ [\alpha_d\lambda_d(k,l) + (1-\alpha_d)|X(k,l)|^2](1-p(k,l))\end{aligned} \tag{6}$$

where $\alpha_d (0 < \alpha_d < 1)$ is a smoothing parameter. $p(k,l)$ denotes the conditional speech presence probability which is related to $\gamma(k,l)$ and $\xi(k,l)$. The *a priori* SNR is estimated as follow:

$$\begin{aligned}\xi(k,l) =& \alpha G^2(k,l-1)\gamma(k,l-1) \\ &+ (1-\alpha)\max\{\gamma(k,l-1)-1, 0\}\end{aligned} \tag{7}$$

where α is a weighting factor that controls the tradeoff between noise reduction and speech distortion [23] and $G(k,l-1)$ denotes the gain function used to recover speech which is related to $\gamma(k,l)$ and $\xi(k,l)$. More details can be found in [22].

As for the traditional frame-level IMCRA approach, the *a priori* SNR and *a posteriori* SNR are defined at frame level. This approach pays more attention on the update through frame instead of global information so that in the process of entering the speech frame from noise frame, the SNR will change dramatically, which may affect the performance of the method.

2.3. Multi-level adaptive noise and speech estimation

According to description of Section 2.2, $\gamma(k,l)$ is related to $\gamma(k,l-1)$ and $\xi(k,l-1)$ while $\xi(k,l)$ depends on $G(k,l-1)$ and $\gamma(k,l-1)$. Therefore, the gain function, the *a priori* SNR, and *a posterior* SNR are recursively coupled between consecutive frames, as shown in Fig. 1. To improve the accuracy of these three estimations in adverse environments, masks estimated by BLSTM-PL are utilized in the estimation procedure. $\hat{M}_{\text{PRM}_1}$ is incorporated to define a new gain function G_{ANSE} because of its strong ability in speech preservation:

$$G_{\text{ANSE}}(k,l-1) = \delta\sqrt{\hat{M}_{\text{PRM}_1}(k,l-1)} + (1-\delta)G(k,l-1) \tag{8}$$

where δ is a weighting factor set to 0.5. We can use G_{ANSE} to recover the STFT of clean speech frame by frame as follow:

$$S(k,l) = G_{\text{ANSE}}(k,l)X(k,l) \tag{9}$$

In Eq. (7), α is a constant controlling the tradeoff between noise reduction and speech distortion. However, in a real life scenario, the target speech is usually destroyed by kinds of stationary and non-stationary background noise. So it is very hard for the IMCRA method to track the changes of noise signal in time due to the constant smoothing parameter, especially when the dynamic range of noise is large. In [17], the authors find that the PL model can achieve better performance on controlling the tradeoff between noise reduction and speech distortion, and in terms of ASR performance, the output $\hat{M}_{\text{PRM}_1}(k,l)$ is better than the other outputs, $\hat{M}_{\text{PRM}_2}(k,l)$ and $\hat{M}_{\text{PRM}_3}(k,l)$. The $\hat{M}_{\text{PRM}_1}(k,l)$ can provide accurate voice activity detection (VAD) information and extend the α to an adaptive smoothing factor at each time-frequency bin. The $priori$ SNR $\xi(k,l)$ at the l^{th} frame is calculated as follow:

$$\begin{aligned}\xi(k,l) =& \alpha(k,l)G_{\text{ANSE}}(k,l-1)^2\gamma(k,l-1) \\ &+ (1-\alpha(k,l))\max\{\gamma(k,l-1)-1, 0\}\end{aligned} \tag{10}$$

$$\alpha(k,l) = (1-\hat{M}_{\text{PRM}_1}(k,l))\alpha_{\max} + \hat{M}_{\text{PRM}_1}(k,l)\alpha_{\min} \tag{11}$$

where $\alpha(k,l)$ is the adaptive smoothing factor. $\alpha_{\max}$ and $\alpha_{\min}$ are the maximum and minimum values of the dynamic range of α. $\hat{M}_{\text{PRM}_1}$ is combined in the formula because learning speech of lower SNRs makes the mask perform better in speech preservation. The parameter α can be adjusted automatically so that PL-ANSE has the ability to control the ability of noise reduction according to the estimated mask.

For Eq. (6), the conditional speech presence probability, $p(k,l)$, is used to control the update of noise estimation. In this paper, the outputs of $\hat{M}_{\text{PRM}_2}$ and $\hat{M}_{\text{PRM}_3}$ are also adopted to improve the noise estimation as follow:

$$\begin{aligned}\lambda_d(k,l+1) &= \lambda_d(k,l)\hat{p}(k,l) \\ &+ [\alpha_d\lambda_d(k,l) + (1-\alpha_d)|X(k,l)|^2](1-\hat{p}(k,l))\end{aligned} \tag{12}$$

where $\alpha_d (0 < \alpha_d < 1)$ is a smoothing parameter. $\hat{p}(k,l)$ denotes a variable consisting of $\hat{M}_{\text{PRM}_2}$, $\hat{M}_{\text{PRM}_3}$ and the conditional speech presence probability $p(k,l)$.

$$\hat{p}(k,l) = bp(k,l) + \frac{1-b}{2}(\sqrt{\hat{M}_{\text{PRM}_2}} + \sqrt{\hat{M}_{\text{PRM}_3}}) \tag{13}$$

where b is a weighting factor whose value ranges from 0 to 1. $\hat{M}_{\text{PRM}_2}$ and $\hat{M}_{\text{PRM}_3}$ are combined with $p(k,l)$ because the meaning of clean or high-SNR masks has common properties with speech presence probability. The mask with a large value also indicates that the possibility of speech presence is high. Through combining accurate masks, noise spectrum will be better estimated.

3. EXPERIMENTS AND RESULT ANALYSIS

3.1. Data corpus

CHiME-4 noise is chosen as our noise database which consists of noise in four conditions: cafe (CAF), street junction (STR), public transport (BUS), and pedestrian area (PED) [24]. Clean speech is derived from the WSJ0 corpus. 7138 utterances (about 12 hours of reading style speech) from 83 speakers, denoted as SI-84 training set, are corrupted with the CHiME-4 noise at three SNR levels (-5dB, 0dB, 5dB) to build a 36-hour training set, consisting of pairs of clean and noisy utterances. We present the experimental evaluation of our framework in the CHiME-4 real test dataset. The real speech in this dataset were collected from talkers reading the same sentences from the WSJ0 corpus in the four conditions. Then we test the effect of our proposed method on the development and test sets of real and simulated speech.

3.2. Implementation details

For front-end configurations, speech waveform is sampled at 16 kHz. The corresponding frame length and frame shift are set to 512 samples and 256 samples, respectively. The short-term Fourier analysis is used to compute the DFT of each overlapping windowed frame. We used pytorch to train the PL network. The model was trained for 30 epochs. The learning rate was initialized as 0.25 and then decreased by 20% after every 6 epochs. The batch size was 8. The loss weight parameter η_m of each layer was 1. In PL-ANSE approach, $\alpha_{\min}$ and $\alpha_{\max}$ were set to 0.7 and 0.96, respectively. The other parameters were set according to [22].

For ASR system, which officially provided in [24] was adopted to evaluate our approach without acoustic models retraining. The acoustic model is a DNN-HMM discriminatively trained with the sMBR criterion [25]. The language models are 5-gram with Kneser-Ney (KN) smoothing for the first-pass decoding [26] and the simple RNN-based language model for rescoring.

3.3. Experiments on PL-ANSE

Table 1. WER (%) comparison of different approaches on the real test set.

Enhancement	BUS	CAF	PED	STR	AVG
Noisy	36.55	24.73	19.92	14.16	23.84
IMCRA	33.91	26.62	21.88	13.17	23.89
PRM(T1)	32.44	21.20	16.95	11.43	20.50
PRM(T2)	37.02	30.18	23.08	13.71	25.99
IRM(T3)	39.41	33.02	24.83	14.05	27.82
ISPP(T1)	29.52	23.87	18.82	11.84	21.01
ISPP(T3)	29.76	24.93	20.01	11.71	21.60
PL-ANSE	**27.99**	**20.56**	**15.81**	**9.92**	**18.57**

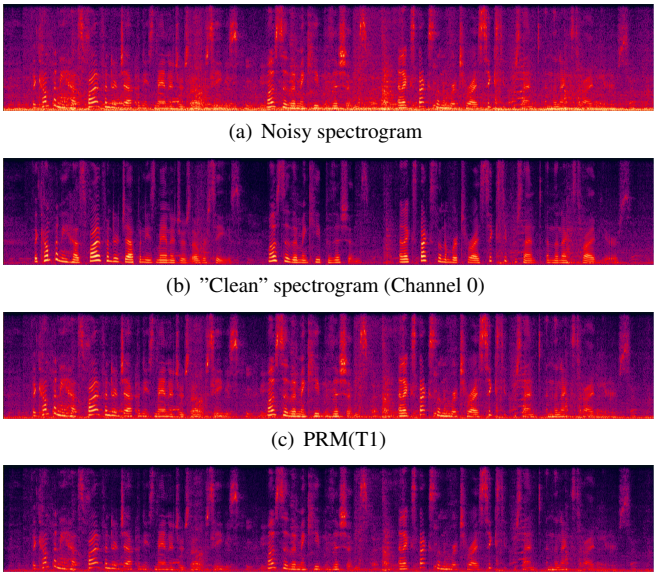
Table 1 shows the comparison of different methods for word error rate (WER) on the CHiME-4 real test set. In the first block from top to bottom, "Noisy" refers to the case where the real noisy speech is directly fed into the recognition system. "IMCRA" denotes that the enhanced speech is obtained by the gain function calculated by the traditional speech enhancement algorithm IMCRA. In the second block, "PRM(T1)" and "PRM(T2)" represent the output masks of the first two target layers of the PL structure. While "IRM(T3)" denotes the final output of PL network. The average WER of 4 conditions of "Noisy" is 23.84%. For the three targets of PL, the PRM of the first target layer "PRM(T1)" achieved the lowest WER of 20.50% while the other two targets "PRM(T2)" and "IRM(T3)" both destroyed the ASR performance when compared to unprocessed "Noisy" speech. This is because the latter two methods damage the spectrum of useful speech while "PRM(T1)" has the most conservative enhancement effect.

In the third block of Table 1, "ISPP(T1)" and "ISPP(T3)" represent merging $\hat{M}_{\text{PRM}_1}$ and $\hat{M}_{\text{IRM}}$ with the gain function in IMCRA according to [21], respectively. "PL-ANSE" denotes PL-ANSE approach we present in Section 2. The WER of "ISPP(T1)" shows that directly combine PRM and gain function of IMCRA does not further improve the ASR performance. But we also find out that in the situation "BUS", there is a big difference between real test data and simu training data, and "ISPP(T1)" performed 2.92% better than "PRM(T1)". The integration of IRM into ISPP will have some improvements compared with "IRM(T3)", but its WER is still higher than "PRM(T1)". For our proposed "PL-ANSE" approach, the ASR performance is obviously better than "PRM(T1)" by utilizing more prior information well. "PL-ANSE" outperforms all other methods in all four cases, and finally achieves a relative WER reduction of 22.1% when compared to "Noisy".

Table 2. Average WER (%) comparison of PL-ANSE and mask-based PL approach on the development and test sets across four environments.

Enhancement	Development set		Test set	
	SimData	RealData	SimData	RealData
Noisy	13.01	11.67	20.88	23.84
PRM(T1)	11.77	10.92	18.05	20.50
PL-ANSE	**11.09**	**9.27**	**17.41**	**18.57**

Table 2 shows the average WER comparison of "PRM(T1)" and "PL-ANSE" on the development and test sets across four environments. The result demonstrates that our proposed approach is more effective on realistic data under adverse environments.



(a) Noisy spectrogram

(b) "Clean" spectrogram (Channel 0)

(c) PRM(T1)

(d) PL-ANSE

Fig. 2. An example of the spectrograms of noisy speech, "clean" speech, "PRM(T1)" and "PL-ANSE".

In Fig. 2, a representative sample utterance from the Real-Data test set is selected to intuitively compare different methods. Fig. 2(a) plots the noisy spectrogram and the speech is severely corrupted with BUS noise. Fig. 2(b) plots the spectrogram from Channel 0 (the close-talking microphone to record the reference "clean" speech). Fig. 2(c) and (d) plot the spectrogram of the first intermediate layer of PL model and our proposed PL-ANSE, respectively. From the spectrogram of "PRM(T1)" we can find out that although "PRM(T1)" is not very powerful in noise reduction, it does not damage the spectrum. On the basis of "PRM(T1)", PL-ANSE obtains stronger noise reduction ability and does a good job in speech preservation at the same time. Based on the above results, we can find that the deep learning-based PRMs can provide accurate prior information to help the conventional speech enhancement estimate the noise and speech spectrums frame by frame.

4. CONCLUSION

In this study, we propose a PL-ANSE approach as a front-end system to improve the performance of ASR systems. First, we design a BLSTM-PL model to estimate PRMs and IRM. Then, we use the PRM of the first intermediate layer to help calculate the gain function and adjust the weighting factor which controls the tradeoff between noise reduction and speech distortion. Next, the other two masks are used to help noise estimation. Finally, the enhanced speech is directly fed into ASR systems to improve its noise robustness. ASR experiment results show that our proposed PL-ANSE achieves a relative WER reduction of 22.1% when compared to that obtained with unprocessed noisy speech on the CHiME-4 test set.

5. ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202. This work was also funded by Tencent.

6. REFERENCES

[1] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition," *Bell Labs Technical Journal*, vol. 62, no. 4, pp. 1035–1074, 2013.

[2] 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.

[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] 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.

[6] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3709–3713.

[7] 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.

[8] 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.

[9] A. Narayanan and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 92–101, 2015.

[10] W. Hartmann, A. Narayanan, E. Fosler-Lussier, and D. Wang, "A direct masking approach to robust asr," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 1993–2005, 2013.

[11] A. Narayanan and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 92–101, 2015.

[12] Szu Jui Chen, Aswin Shanmugam Subramanian, Hainan Xu, and Shinji Watanabe, "Building state-of-the-art distant speech recognition using the chime-4 challenge with a setup of speech enhancement baseline," in *Interspeech 2018*, 2018.

[13] 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.

[14] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 577–581.

[15] T. Gao, J. Du, L. Dai, and C. Lee, "Densely connected progressive learning for lstm-based speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5054–5058.

[16] 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.

[17] N. Zhou, J. Du, Y. Tu, T. Gao, and C. Lee, "A speech enhancement neural network architecture with snr-progressive multi-target learning for robust speech recognition," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 873–877.

[18] McAulay, R., Malpass, and M., "Speech enhancement using a soft-decision noise suppression filter," *Acoustics Speech & Signal Processing IEEE Transactions on*, 1980.

[19] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[20] Y. Tu, I. Tashev, S. Zarar, and C. Lee, "A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2531–2535.

[21] 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.

[22] 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.

[23] O. Cappe, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[24] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon P Barker, and Ricard Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, 2016.

[25] G. Saon and H. Soltau, "A comparison of two optimization techniques for sequence discriminative training of deep neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5567–5571.

[26] 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.