A Speech Enhancement Neural Network Architecture with SNR-Progressive Multi-Target Learning for Robust Speech Recognition

Nan Zhou* and Jun Du* and Yan-Hui Tu* and Tian Gao† and Chin-Hui Lee‡
* University of Science and Technology of China, Hefei, Anhui, P. R. China
E-mail: {zhounan1,tuyanhui}@mail.ustc.edu.cn, jundu@ustc.edu.cn
† iFlytek Research, Hefei, Anhui, P. R. China
‡ Georgia Institute of Technology, Atlanta, Georgia, USA

Abstract—**We present a pre-processing speech enhancement network architecture for noise-robust speech recognition by learning progressive multiple targets (PMTs). PMTs are represented by a series of progressive ratio masks (PRMs) and progressively enhanced log-power spectra (PELPS) targets at various layers based on different signal-to-noise-ratios (SNRs), attempting to make a tradeoff between reduced background noises and increased speech distortions. As a PMT implementation, long short-term memory (LSTM) is adopted at each network layer to progressively learn intermediate dual targets of both PRM and PELPS. Experiments on the CHiME-4 automatic speech recognition (ASR) task, when compared to unprocessed speech using multi-condition trained LSTM-based acoustic models without retraining, show that PRM-only as the learning target can achieve a relative word error rate (WER) reduction of 6.32% (from 27.68% to 25.93%) averaging over the RealData evaluation set, while conventional ideal ration masks severely degrade the ASR performance. Moreover, the proposed LSTM-based PMT network, with the best configuration, outperforms the PRM-only model, with a relative WER reduction of 13.31% (further down to 22.48%) averaging over the same test set.**

Index Terms: progressive ratio mask, progressively enhanced log-power spectra, progressive multi-targets, deep learning based speech enhancement, robust speech recognition

I. Introduction

Recently, hands-free speech interaction with smart phones and artificial intelligence speakers equipped with automatic speech recognition (ASR) capabilities is becoming an essential voice input mode, due to the rapid development of ASR technology [1]–[3]. However, speech signal, corrupted by reverberation and background noise, may degrade ASR system performances, especially in realistic adverse environments [4]. Accordingly, single-channel and multi-channel speech enhancement methods to improve the robustness of ASR systems has attracted quite a bit of research attentions [5]–[9].

Single-channel speech enhancement, especially based on deep learning [9]–[16], has been studied by many researchers to improve ASR performance, but the acoustic models (AMs) under multi-condition training are also often required to be retrained. In [17], a bidirectional long short-term memory (BLSTM) neural network was used to accurately predict ideal ratio mask (IRM) for speech enhancement, denoted as BLSTM-IRM. And the experiments show that BLSTM-IRM was not effective without AM retraining because there exists mismatch between enhanced speech and the training data distributions represented by AMs. Therefore, it is quite challenging for pre-processing approaches to yield performance gains on AMs using multi-condition training without retraining [18]. In this paper we focus our attention on single-channel speech enhancement with a set of progressively learned multiple targets to ease this difficulty.

Based on the above analysis, conventional deep learning based speech enhancement methods usually directly learns the clean spectral features or IRM, given the noisy spectral features, but it is very hard for neural networks to learn this non-linear relationship well especially under mismatched low signal-to-noise ratio (SNR) Conditions. So, in [19]–[21], SNR-based progressive learning for speech enhancement was 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 eventually reaches the clean spectral features. Experiments show that it can improve the performance compared to conventional deep learning models in terms of speech enhancement metrics.

In this study, we propose a novel pre-processing neural network by designing progressive multi-targets (PMTs) based on SNRs to improve ASR performance without AM retraining, which is comprehensively extended from the previous work [19] with new contributions listed as follows: (1) first, a new learning target, namely the progressive ratio mask (PRM), is proposed; (2) the pre-processing neural network with PMTs aims to learn a series of dual targets of both PRM and progressively enhanced log-power spectra (PELPS); and (3) we evaluate on ASR performance rather than the conventional speech enhancement metrics. The tradeoff between background noise reductions and introduced nonlinear distortions can be controlled conveniently by PMTs at different target layers. As an implementation of PMT network, LSTM is adopted at each layer to progressively learn one pair of intermediate targets of PRM and PELPS. The whole network can be optimized in a multi-task learning manner. Experiments

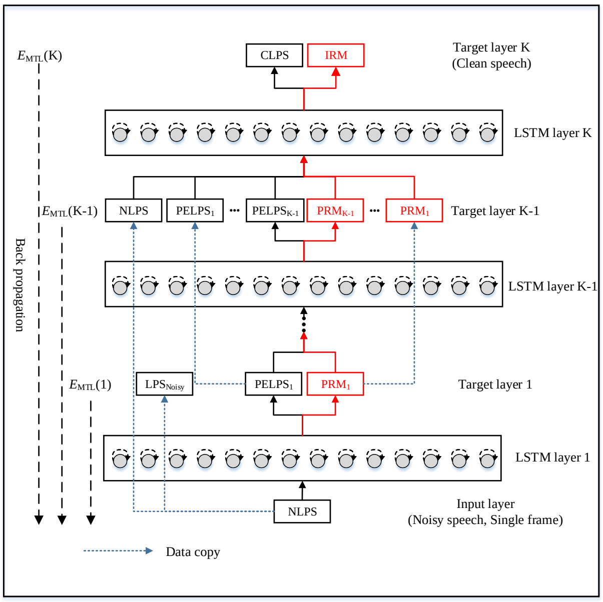


Fig. 1. *An illustration of progressive multi-targets network.*

on the CHiME-4 ASR task with RealData testing sets show that, using the same LSTM architecture, the PRM as the learning target can achieve good word error rate (WER) reductions when compared to unprocessed noisy speech in the setting of multi-condition trained AMs without retraining, with a relative WER reduction of 6.32% averaging over the CHiME-4 RealData evaluation set, while the conventional IRM severely degrades the ASR performance in this case. Moreover, the proposed LSTM-based PMT network with the best setting significantly outperforms the LSTM-based single PRM-only model, with a relative WER reduction of 13.31% averaging over the same test set.

II. Progressive Multi-target Neural Network

A block diagram of the proposed progressive multi-targets network is shown in Figure 1. The input of the whole network is noisy LPS (NLPS) and only the current frame is used without frame expansion. The whole diagram can be divided into two branches with black solid lines and red solid lines. The branch with black solid lines shows the progressive learning network proposed in [19], [20], and only the PELPS is used. The branch with red solid lines demonstrates our proposed progressive ratio masks. The whole PMT network is divided into successively stacking blocks with one LSTM layer and one fully connected layer via dual-target learning per block. The fully connected layer in every block is also referred to as a target layer, which is designed to learn intermediate speech targets with a higher SNR than the targets of previous target layers. Each target layer has dual targets, namely PELPS and PRM.

A. Progressive ratio mask

In [22]–[24], the IRM-based speech enhancement is proposed, and has shown its effectiveness in terms of speech enhancement metrics. However, in [17], the BLSTM-IRM based speech enhancement failed in improving ASR performance without AM retraining, and the reason is that there exists mismatch between enhanced speech distribution and the training data distribution represented by AMs. In [20], the authors also demonstrate that it is hard for the deep-learning-based regression model to directly learn the relationship between the noisy LPS features and clean LPS features. In this study, we propose the PRM target to make a tradeoff between the noise reduction and speech distortion. And the PRM is defined as

$$z^{\text{PRM}}(t,f) = \frac{S(t,f) + N_{\text{T}}(t,f)}{S(t,f) + N_{\text{I}}(t,f)} \tag{1}$$

where $S(t,f)$ represents the power spectrum of the speech signal at the time-frequency (T-F) unit (t,f), $N_{\text{T}}(t,f)$ and $N_{\text{I}}(t,f)$ represent the power spectrum of the noise in one PRM target and input signals at the T-F unit (t,f), respectively. When the numerator of Eq. (1) becomes the power spectrum of the clean speech signal, $N_{\text{T}}(t,f)$ is zero and $z^{\text{PRM}}(t,f)$ is regressed to the traditional IRM $z^{\text{IRM}}(t,f)$. We should emphasize that PRM can be used not only in our PMT network illustrated by Figure 1, but also as a single learning target just as the IRM using the conventional deep architectures. More details will be discussed in Section III-C.

B. Progressive multi-target learning

In speech enhancement domain, multi-task learning methods such as combining spectrum mapping and mask-based tasks together not only reduce the generalization error but also extract complementary information from multiple tasks. In [25], the authors demonstrate that combining spectrum mapping and mask-based tasks together achieves further improvements in speech enhancement quality, especially in speech intelligibility measures. Therefore we propose the progressive multi-target learning network as illustrated in Figure 1. Every block functions as a sequence learning component to estimate its targets, while the input and the estimations of all previous intermediate targets are spliced together and fed to the LSTM layer in the block, forming the similar structure to DenseNet [26]. Each target layer defines two equally important targets, namely PELPS and PRM.

To implement multi-task learning in PMT network, each target layer is designed to predict both PELPS and PRM targets. The multi-task error between the output of target layer k $(1 \leq k \leq K)$ and its ground-truth label is

$$\begin{aligned} E_{\text{MTL}}(k) \quad = \quad & \sum_{t,f} \left[(\hat{z}^{\text{PELPS}}(k,t,f) - z^{\text{PELPS}}(k,t,f))^2 \right. \\ & \left. + (\hat{z}^{\text{PRM}}(k,t,f) - z^{\text{PRM}}(k,t,f))^2 \right], \end{aligned} \tag{2}$$

where $\hat{z}^{\text{PELPS}}(k,t,f)$ and $z^{\text{PELPS}}(k,t,f)$ are predicted and ground-truth PELPS features of the k^{th} target layer, while $\hat{z}^{\text{PRM}}(k,t,f)$ and $z^{\text{PRM}}(k,t,f)$ are predicted and ground-truth PRM features of the k^{th} target layer. Both $\hat{z}^{\text{PELPS}}(k,t,f)$ and $\hat{z}^{\text{PRM}}(k,t,f)$ are nonlinear functions of PELPS and PRM in preceding target layers. $z^{\text{PELPS}}(k,t,f)$ and $z^{\text{PRM}}(k,t,f)$ can be easily calculated with a predefined SNR gain of target layer k. K is the number of target layers. Please note that PELPS and PRM of target layer K correspond to clean LPS (CLPS) features and IRM, respectively. The errors of all target layers are computed in the mean squared error (MSE) sense, and added together to optimize the trainable parameters.

In the enhancement stage, every target layer has two outputs, predicting the corresponding PELPS and PRM. We can use either of them as the preprocessed result for ASR system. To combine these two targets for further improving ASR performance, an ensemble method via a simple average is adopted:

$$\begin{aligned}&\hat{z}^{\text{Fusion}}(k,t,f) = \\ &\frac{1}{2}\left[\hat{z}^{\text{PELPS}}(k,t,f) + \log \hat{z}^{\text{PRM}}(k,t,f) + x^{\text{LPS}}(t,f)\right]\end{aligned} \tag{3}$$

where $\hat{z}^{\text{Fusion}}(k,t,f)$ is the fusion result of two outputs at T-F unit (t,f) in the target layer k, $x^{\text{LPS}}(t,f)$ is the input noisy LPS feature at T-F unit (t,f). When using our PMT network as a preprocessor for a specific ASR system, one critical issue is how to select the optimal target from all target K layers. One simple way is to determine it in terms of a lowest WER on a development set and apply it to the test/evaluation set.

III. Experiments and Result Analysis

A. Speech enhancement systems

CHiME-4 noises (BUS, CAF, PED and STR) [27] were chosen as the noise database to match the training and test conditions for our AM. Clean speech was derived from the WSJ0 corpus with 7138 utterances (about 12 hours of reading style speech) by 83 speakers, denoted as SI-84 training set. Clean utterances were corrupted with the above mentioned CHiME-4 noises at three SNR levels (-5dB, 0dB and 5dB) to build our training set by the data simulation method [27].

Speech waveform was sampled at 16 kHz, and the corresponding frame length was set to 512 samples (or 32 msec) with a frame shift of 256 samples. A short-time Fourier analysis was used to compute the DFT of each overlapping windowed frame. Then the 257-dimensional LPS features normalized by global mean and variance were used to train neural networks using Microsoft Computational Network Toolkit (CNTK) [28]. The network, as depicted, is composed of K stacking blocks, where K is usually 3, 5 or 7, each with a different design of SNR gains as described in Table II. The enhancement models were trained using Adam optimizer for 20 epochs.

B. Speech recognition systems

The CHiME-4 challenge baseline AM [27], namely a DNN-HMM with 7 layers and 2048 neurons per layer, was used as the multi-condition trained AM without retraining in this study. The channel-5 noisy training data of CHiME-4 was used to train the AM while the development and evaluation sets of 1-channel track of CHiME-4 real data were used in the recognition stage. The acoustic features to train the AM were based on feature-space maximum likelihood linear regression (fMLLR) transformation, and the AM was optimized by sequence discriminative training. 3-gram language model was used and more details of ASR system could refer to [27].

C. Experiments on progressive ratio masks (PRMs)

Table I shows that comparison of different SE models on the development and test sets of RealData. For the first block from the top, "Noisy" denotes original speech randomly selected from channel 1-6 (except channel 2), namely single-channel case. The second block from the top shows that enhanced speech is obtained by the estimated IRM and PRM, using the same 257-1024-1024-257 LSTM architecture. PRM(T1) and PRM(T2) denote that the PRM is calculated with +10dB SNR gain and +20dB SNR gain, respectively.

As a summary, the IRM estimated by LSTM is not effective for improving ASR performance. For example, the WERs of "Noisy" are 15.68% and 27.67%, while the WERs of "LSTM-IRM" are 20.76% and 35.09% on the development and evaluation sets in average, respectively. Next, for "LSTM-PRM(T1)", the proposed PRM estimated by LSTM can improve the ASR performance directly when comparedg to the "Noisy" row with a relative WER reduction of 3.07% and 6.32% (from 27.68% shown in the rightmost column of the top row in Table 1 to 25.93% shown in the rightmost column of the second row) on the development and evaluation sets in average, respectively. Finally, for "LSTM-PRM(T2)", it destroyed the ASR performance when compared to unprocessed "Noisy" speech, but the performance of "LSTM-PRM(T2)" is better than that of "LSTM-IRM".

D. Experiments on progressive multi-targets (PMTs)

Based on the above analysis, we can find that the proposed PRMs with different SNR gains have different influences on the tradeoff between noise reduction and speech distortion. So, the PMT model is proposed to estimate different PRMs and PELPSs at different target layers.

The bottom three blocks of Table I show that enhanced speech is obtained by the estimated PRM, PELPS, and fusion method using the PMT model with 3 target layers. PMT-3T denotes the LSTM-based progressive multi-target model with 3 target layers are used. "T1", "T2" and "T3" denote the output of target layer 1, 2 and 3, and also corresponding to the PRM(T1), PRM(T2) and IRM, respectively. PRM, PELPS and Fusion denote enhanced speech is obtained by PRM output, PELPS output and the ensemble of the two outputs of PMT-3T.

First, "PMT-3T-T1-PRM", the intermediate target with +10dB SNR gain, can obtain the best performance at all situations. And it also can improve the ASR performance when compared to that of LSTM-PRM(T1) (shown as 27.68%) at the rightmost column in the top row of the second block of Table I, with a relative WER reduction of 8.71% and 10.37% on RealData development and evaluation sets, respectively. Furthermore, fusion of PRM and PELPS can further improve the ASR performance when the performance of PELPS and PRM is comparable as shown in T1 and T2 of Table I.

TABLE I
OVERALL WER (%) COMPARISON OF DIFFERENT SE MODELS ON THE DEVELOPMENT AND TEST SETS OF REALDATA.

	Dev Set (Real)					**Eval Set (Real)**				
SE model	BUS	CAF	PED	STR	AVG	BUS	CAF	PED	STR	AVG
Noisy	18.96	15.29	9.87	14.61	14.68	40.42	28.97	24.27	17.09	27.68
LSTM-PRM(T1)	**16.40**	**17.05**	**10.43**	**13.05**	**14.23**	**37.26**	**29.32**	**22.10**	**15.04**	**25.93**
LSTM-PRM(T2)	17.52	20.30	11.77	15.89	16.37	38.45	35.22	25.38	17.56	29.15
LSTM-IRM(T3)	18.53	25.52	13.10	17.56	18.68	42.64	39.69	28.51	19.54	32.59
PMT-3T-T1-PRM	**15.44**	**15.10**	**9.29**	**12.14**	**12.99**	**33.03**	**25.46**	**19.41**	**15.04**	**23.24**
PMT-3T-T2-PRM	17.30	19.78	11.90	14.22	15.80	34.74	33.54	23.71	16.78	27.19
PMT-3T-T3-PRM	18.17	22.06	12.17	15.51	16.98	36.90	34.90	24.81	17.20	28.32
PMT-3T-T1-PELPS	**15.19**	**16.31**	**9.57**	**12.11**	**13.30**	**33.24**	**26.47**	**19.88**	**14.42**	**23.50**
PMT-3T-T2-PELPS	16.88	20.36	11.19	13.87	15.57	36.28	31.05	22.64	16.20	26.54
PMT-3T-T3-PELPS	22.69	36.47	18.42	24.20	24.45	44.72	50.58	37.46	23.38	39.03
PMT-3T-T1-Fusion	**15.49**	**14.99**	**9.51**	**11.61**	**12.90**	**32.23**	**24.45**	**19.71**	**14.53**	**22.73**
PMT-3T-T2-Fusion	16.98	19.50	11.26	13.94	15.42	36.12	30.59	23.19	16.01	26.47
PMT-3T-T3-Fusion	20.64	27.02	14.16	18.61	20.11	40.27	40.16	31.24	19.85	32.88

TABLE II
SNR GAIN CONFIGURATIONS FOR THE TARGET LAYERS.

K	SNR gains for each target layer
3	10dB (Target 1-2)
5	5dB (Target 1-4)
7	2.5dB (Target 1-4), 5dB (Target 5-6)

Based on the above analysis, we can find the target of PELPS and PRM with +10dB SNR gain is optimal for the multi-condition trained AM we utilized in all ASR experiments. Therefore, we increase from 3 to 5 tagets (5T) and 7 targets (7T) and directly summarize the results of optimal target layer output of PMT-3T, PMT-5T and PMT-7T models in Table. III. First, we can find that the fusion performance can obtain stable improvements in all three models. Second, the fusion performance of PMT-7T is better than that of PMT-3T and PMT-5T. Finally, the proposed LSTM-based PMT network, with the best configuration, outperforms the LSTM-based single PRM-only model, with a relative WER reduction of 13.31% (further down from 25.93% at the rightmost column of the top row in Table 1 to 22.48% at the rightmost column of the bottom row in Table 3) averaging over the RealData evaluation set.

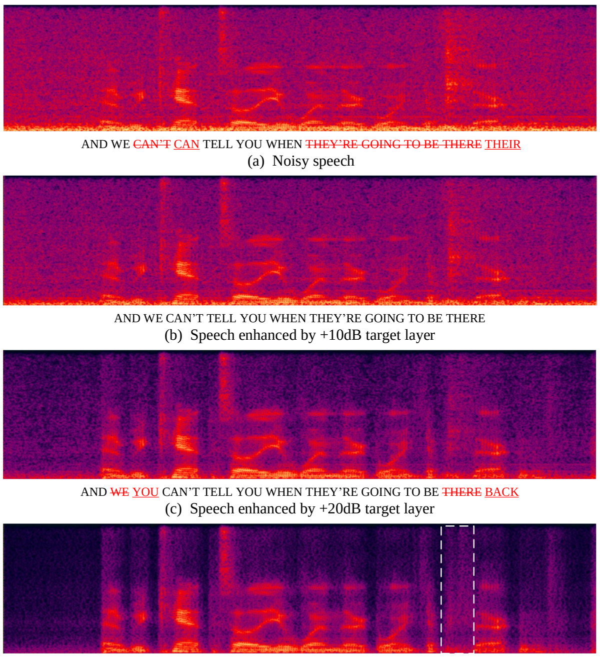
TABLE III
WER(%) RESULTS OF OPTIMAL TARGET LAYER OUTPUT OF PMT-3T, PMT-5T AND PMT-7T MODELS ON EVALUATION SET OF REALDATA.

SE model	BUS	CAF	PED	STR	AVG
PMT-3T	32.23	24.45	19.71	14.53	22.73
PMT-5T	32.40	**24.09**	19.71	14.20	22.60
PMT-7T	**32.15**	24.13	**19.56**	**14.05**	**22.48**

In Figure 2, we selected a representative sample utterance from the RealData evaluation set for a further analysis. The white box in the bottom spectrogram emphasizes the region that are severely corrupted with speech distortions, resulting in vanished speech information. From top to bottom, the SNR targets were gradually increased and the background noises were reduced. So the estimated clean speech target might show good listening quality. However, with increased SNR, the non-linear distortions introduced by the enhancement models were also increased. For ASR, both high-level background noises and speech distortions could lead to substitution, insertion and deletion errors marked red. It seemed that the target layer with +10dB SNR gain in the second spectrogram from the top made the best tradeoff, yielding a totally correct recognition result for this utterance.

IV. CONCLUSION

In this study, we extend our previous work of progressive neural network based speech enhancement to learning multiple-SNR targets and using it as a pre-processor for noise-robust ASR. We investigate designing intermediate enhancement targets so that the pre-processor can be directly used at the recognition stage without retraining AMs. First, a new learning target, namely the PRM, is proposed. Second, the pre-processing neural network with PMTs aims at learning a series of dual targets of both PRM and PELPS. Third, we evaluate on ASR performance rather than the conventional speech enhancement metrics. The tradeoff between reduced background noises and increased speech distortions can be controlled conveniently by PMTs at different target layers. Experiments on the CHiME-4 ASR task show that the only PRM as the learning target can achieve good WER reductions when compare to unprocessed speech using multi-condition trained AMs without retraining, with a relative WER reduction of 6.32% averaging over the CHiME-4 RealData evaluation set. Furthermore the proposed LSTM-based PMT network with the best configurations outperforms the LSTM-based single PRM-only model, with a further relative WER reduction of 13.31% averaging over the same test set. In the future, we will verify the effectiveness of our proposed PMT model on



AND WE ~~CAN'T~~ CAN TELL YOU WHEN ~~THEY'RE GOING TO BE THERE~~ THEIR

(a) Noisy speech

AND WE CAN'T TELL YOU WHEN THEY'RE GOING TO BE THERE

(b) Speech enhanced by +10dB target layer

AND ~~WE~~ YOU CAN'T TELL YOU WHEN THEY'RE GOING TO BE ~~THERE~~ BACK

(c) Speech enhanced by +20dB target layer

AND ~~WE~~ YOU CAN'T TELL YOU WHEN ~~THEY'RE GOING TO BE THERE~~ THEIR POINT OF FACT

(d) Speech enhanced by clean target layer

Fig. 2. *An example of 3 outputs from 7-target PMTL model with the spectrograms and transcriptions by a multi-condition trained acoustic model: (a) noisy speech, (b) speech enhanced by +10dB target layer, (c) speech enhanced by +20dB target layer, (d) speech enhanced by clean target layer.*

the CHiME-5 challenge.

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)*. IEEE, 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)*. IEEE, 2018, pp. 5934–5938.

[3] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[4] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust automatic speech recognition: a bridge to practical applications*. Academic Press, 2015.

[5] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5745–5749.

[6] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5210–5214.

[7] J. Li, Y. Huang, and Y. Gong, "Improved cepstra minimum-mean-square-error noise reduction algorithm for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4865–4869.

[8] Y. Tu, J. Du, L. Sun, F. Ma, and C.-H. Lee, "On design of robust deep models for chime-4 multi-channel speech recognition with multiple configurations of array microphones." in *INTERSPEECH*, 2017, pp. 394–398.

[9] 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)*. IEEE, 2018, pp. 5024–5028.

[10] M. Delcroix, Y. Kubo, T. Nakatani, and A. Nakamura, "Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling?" in *Interspeech*, 2013, pp. 2992–2996.

[11] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[12] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4375–4379.

[13] 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.

[14] T. Yoshioka and M. J. Gales, "Environmentally robust ASR front-end for deep neural network acoustic models," *Computer Speech & Language*, vol. 31, no. 1, pp. 65–86, 2015.

[15] K. Han, Y. He, D. Bagchi, E. Fosler-Lussier, and D. Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[16] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[17] S.-J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the chime-4 challenge with a setup of speech enhancement baseline," in *Interspeech*, 2018.

[18] H. Tang, W.-N. Hsu, F. Grondin, and J. Glass, "A study of enhancement, augmentation, and autoencoder methods for domain adaptation in distant speech recognition," *arXiv preprint arXiv:1806.04841*, 2018.

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

[20] ——, "SNR-based progressive learning of deep neural network for speech enhancement." in *INTERSPEECH*, 2016, pp. 3713–3717.

[21] L. Sun, J. Du, T. Gao, Y.-D. Lu, Y. Tsao, C.-H. Lee, and N. Ryant, "A novel LSTM-based speech preprocessor for speaker diarization in realistic mismatch conditions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5234–5238.

[22] 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.

[23] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7092–7096.

[24] ——, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 826–835, 2014.

[25] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017*. IEEE, 2017, pp. 136–140.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[27] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[28] D. Yu, A. Eversole, M. Seltzer, K. Yao, Z. Huang, B. Guenter, O. Kuchaiev, Y. Zhang, F. Seide, H. Wang *et al.*, "An introduction to computational networks and the computational network toolkit," *Microsoft Technical Report MSR-TR-2014–112*, 2014.