DENSELY CONNECTED PROGRESSIVE LEARNING FOR LSTM-BASED SPEECH ENHANCEMENT

Tian Gao[1]*, Jun Du*[1]*, Li-Rong Dai*[1]*, Chin-Hui Lee*[2]

[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, Anhui, China
[2]Georgia Institute of Technology, Atlanta, Georgia, USA

gtian09@mail.ustc.edu.cn, {jundu, lrdai}@ustc.edu.cn, chl@ece.gatech.edu

ABSTRACT

Recently, we proposed a novel progressive learning (PL) framework for deep neural network (DNN) based speech enhancement to improve the performance in low signal-to-noise ratio (SNR) environments. In this study, several new contributions are made to this framework. First, the advanced long short-term memory (LSTM) architecture is adopted to achieve better results, namely LSTM-PL, where each LSTM layer is guided to explicitly learn an intermediate target with a specific SNR gain. However, we observe that the performance of LSTM-PL architecture is easily degraded by increasing the number of intermediate targets due to the possible information loss when involving more target layers. Accordingly, we propose densely connected progressive learning in which the input and the estimations of intermediate targets are spliced together to learn the next target. This new structure can fully utilize the rich set of information from the multiple learning targets and alleviate the information loss problem. Experimental results demonstrate that the dense structure with deeper LSTM layers can yield significant gains of speech intelligibility measure for all noise types and levels. Moreover, the post-processing with more targets tends to achieve better performance.

Index Terms— Progressive learning, long short-term memory, dense structure, post-processing, speech enhancement

1. INTRODUCTION

Speech enhancement is an important front-end of speech processing systems aimed at improving speech quality and intelligibility in the presence of an interfering noise signal. Background noise can affect the performance of speech communication, hearing aids, speech recognition and speaker recognition [1]. Numerous algorithms have been proposed over the past several decades to solve this problem. The conventional algorithms include spectral subtraction [2], Wiener filtering [3], minimum mean squared error (MMSE) estimation [4] and optimally-modified log-spectral amplitude (OM-LSA) speech estimator [5]. Spectral subtraction is one of the first algorithms proposed for noise reduction. However, the resulting enhanced speech often suffers from an annoying artifact called musical noise. OM-LSA utilizes a minima controlled recursive averaging (MCRA) noise estimation [6] approach to avoid the musical noise. One limitation of the conventional speech enhancement algorithms is that they can't improve speech intelligibility effectively [7]. In addition to focus on amplitude, some phase-aware speech enhancement methods were investigated in [8, 9]. For learning based methods, nonnegative matrix factorization (NMF) was investigated in the form of supervised and unsupervised for speech enhancement [10, 11]. The basic idea is to decompose the noisy speech data into bases and weights matrices for the speech and noise, respectively.

Speech enhancement in recent years, with the introduction of deep learning, has made great progress. The supervised deep learning approaches have been investigated from the aspects of learning target, neural network structure, input feature, etc. Xu *et, al.* [12, 13] proposed a deep neural network (DNN) based regression framework to predict the clean log-power spectra (LPS) features [14] from noisy LPS features. In [15], masking techniques were used to make classification on time-frequency (T-F) units for speech enhancement. In addition to the direct prediction of mask, Huang *et, al.* [16] investigated joint optimization of masking functions and neural networks with an extra masking layer. More complex neural network structures, such as long short-term memory (LSTM) network [17] and convolutional neural network (CNN) were investigated in [18, 19, 20]. For the input of neural network, Fu *et, al.* [21] has investigated the time domain waveform by using fully CNN.

However, the challenges of speech enhancement in low signal-to-noise ratio (SNR) still remain. Focus on this challenge, a joint framework combining speech enhancement with voice activity detection (VAD) was proposed in [22] to increase the speech intelligibility in low SNR environments. Meanwhile, multi-task learning (MTL) has also been adopted in speech enhancement. In [23], a multi-objective framework was proposed to improve the generalization capability of regression DNN. Based on MTL method, Jiang *et, al.* [24] adopted DNN-based speech denoising with ideal binary mask (IBM) as the targets at different time-frequency scales simultaneously and collaboratively. Another notable machine learning strategy is the curriculum learning [25] originated from cognitive science. Inspired by curriculum learning, we proposed a novel progressive learning (PL) framework [26] to improve the performance of DNN-based speech enhancement in low SNR environments.

In this paper, we continue to study the progressive learning with advanced LSTM network (LSTM-PL), which has been verified more suitable for sequential speech processing tasks [27, 28, 18, 19]. According to the idea of progressive learning, each hidden layer of the LSTM network is guided to learn an intermediate target with a specific SNR gain explicitly. The subproblem solving in each stage can boost the subsequent learning of the next stage. However, we observe that the performance of LSTM-PL architecture is easily degraded by increasing the number of intermediate targets due to the possible information loss when involving more target layers. In order to alleviate this problem and make full use of the rich set of information from the multiple learning targets, we propose densely connected progressive learning in which the input and the estimations

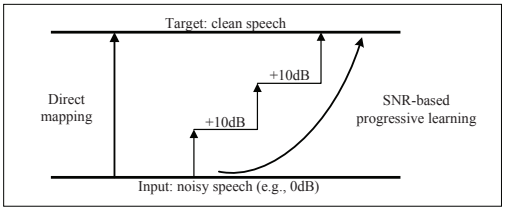

Fig. 1. Progressive learning for speech enhancement [26].

of intermediate target are spliced together to learn next target. Experimental results demonstrate that densely connected progressive learning can make the whole network deeper and yield better speech intelligibility. Moreover, when combined with the multi-target fusion, the proposed approach can be further improved.

2. REVIEW OF DNN-BASED PROGRESSIVE LEARNING

Curriculum learning is related to MTL where the initial tasks are boosted to guide the learner for the better achievement on the final task. However the motivation of MTL is to improve the generalization of the target task by leveraging on other tasks. Inspired by curriculum learning, SNR-based progressive learning was proposed in [26] for DNN-based speech enhancement, as shown in Fig. 1. The basic idea is to start small, learn easier aspects of the task or easier sub-tasks, and then gradually increase the difficulty level. Specific to DNN-based speech enhancement, the direct mapping process from noisy speech to clean speech in the conventional DNN training is decomposed into multiple stages with an SNR gain achieved in each stage. The SNR gains in each stage can boost the subsequent learning of the next stage. For example, if the input SNR of noisy speech is 0dB, the learning target of baseline system is clean speech. And for progressive learning, two new intermediate learning targets (10dB and 20dB speech) will be inserted into the neural network training.

3. LSTM-BASED PROGRESSIVE LEARNING

3.1. LSTM Architecture

In the training of DNN, the important temporal information is only considered via frame expansion. To model time sequences, recurrent neural networks (RNN) seem to have a congenital advantage by using recursive structures between the previous frames and the current frame to capture the long-term contextual information. However, the conventional RNN can not hold information for a long period and the optimization of RNN parameters via the back propagation through time (BPTT) faces the problem of the vanishing and exploding gradients [29]. The problems can be well alleviated by the invention of LSTM [17] which introduces the concepts of memory cell and a series of gates to dynamically control the information flow. Fig. 2 illustrates a single LSTM memory cell. The composite LSTM cell is implemented as follows:

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i) \tag{1}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f) \tag{2}$$

$$\mathbf{c}_t = \mathbf{f}_t \otimes \mathbf{c}_{t-1} + \mathbf{i}_t \otimes \tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) \tag{3}$$

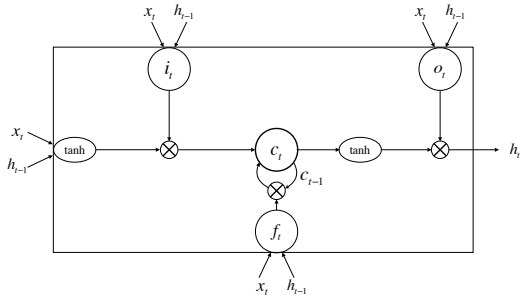

Fig. 2. An illustration of the LSTM cell.

$$\mathbf{o}_t = \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o) \tag{4}$$

$$\mathbf{h}_t = \mathbf{o}_t \otimes \tanh(\mathbf{c}_t) \tag{5}$$

where t is the frame index, σ is the logistic sigmoid function, and $\mathbf{i}$, $\mathbf{f}$, $\mathbf{o}$ and $\mathbf{c}$ are respectively the input gate, forget gate, output gate and cell activation vectors, all of which are the same size as the hidden vector $\mathbf{h}$. $\otimes$ denotes element-wise multiplication. $\mathbf{W}$ and $\mathbf{b}$ represent the weight matrices and bias vectors from the cell to gate, respectively.

3.2. Densely Connected Progressive Learning

In [26], progressive learning has been applied successfully to DNN network. In this study, the LSTM-based densely connected progressive learning is illustrated in Fig. 3 (3 learning targets are defined in this figure). All the target layers are designed to learn intermediate speech with higher SNRs or clean speech. For the input and multiple targets, LSTM layers are used to link between each other. This stacking style network can learn multiple targets progressively and efficiently. In order to make full use of the rich set of information from the multiple learning targets, we update the progressive learning in [26] to densely connected progressive learning in which the input and the estimations of intermediate target are spliced together to learn next target. Then, a weighted MMSE criterion in terms of MTL is designed to optimize all network parameters randomly initialized with K target layers as follows:

$$E = \sum_{k=1}^{K} \alpha_k E_k \tag{6}$$

$$E_k = \frac{1}{N}\sum_{n=1}^{N} \|\mathcal{F}_k(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k) - \mathbf{x}_n^k\|_2^2 \tag{7}$$

where $\hat{\mathbf{x}}_n^k$ and $\mathbf{x}_n^k$ are the n^{th} D-dimensional vectors of estimated and reference target LPS feature vectors for k^{th} target layer, respectively ($k > 0$), with N representing the mini-batch size. $\hat{\mathbf{x}}_n^0$ denotes the n^{th} D-dimensional vector of input noisy LPS features. $\mathcal{F}_k(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, ..., \hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k)$ is the neural network function for k^{th} target with the dense structure using the previously learned intermediate targets from $\hat{\mathbf{x}}_n^0$ to $\hat{\mathbf{x}}_n^{k-1}$, and $\mathbf{\Lambda}_k$ represents the parameter set of the weight matrices and bias vectors before k^{th} target layer, which are optimized in the manner of BPTT with gradient descent.

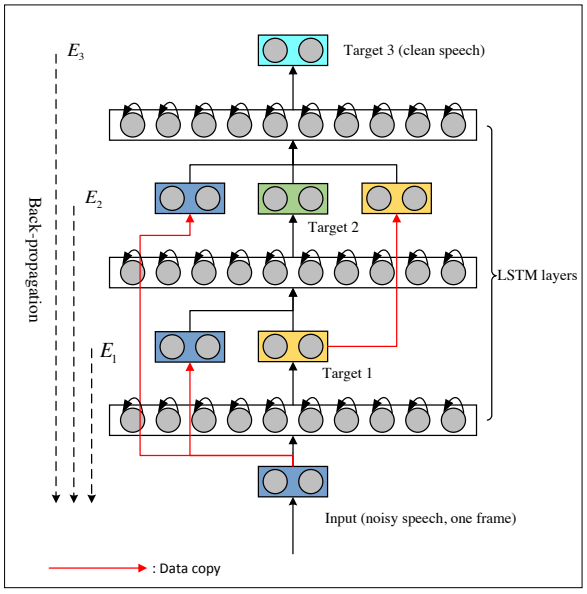

Fig. 3. Densely connected progressive learning for LSTM-based speech enhancement (using three targets as an example).

3.3. Post-processing

One advantage of progressive learning is that there are more than one estimated target from the network. The estimated LPS features of different targets can provide rich information for post-processing. In the testing stage, the estimations of multiple targets are averaged to further improve the overall performance as implemented in [26].

4. EXPERIMENTS AND RESULT ANALYSIS

115 noise types used in [26] are chosen as our noise database. Clean speech is derived from the WSJ0 corpus [30]. 7138 utterances (about 12 hours of reading style speech) from 83 speakers, denoted as SI-84 training set, are corrupted with the above mentioned 115 noise types at three SNR levels (-5dB, 0dB and 5dB) to build a 36-hour training set, consisting of pairs of clean and noisy utterances. The 330 utterances from 12 other speakers, namely the Nov92 WSJ evaluation set, are used to construct the test set for each combination of noise types and SNR levels (-5dB, 0dB, 5dB, 10dB). Six unseen noises from the NOISEX-92 corpus [31], namely babble, factory1, factory2, destroyer engine, m109 and white are adopted for testing. Short-time objective intelligibility (STOI) [32] and source-to-distortion ratio (SDR) [33] are used to assess the intelligibility and SNR of the enhanced speech.

As for the front-end, speech waveform is sampled at 16 kHz, and the corresponding frame length is set to 512 samples (or 32 msec) with a frame shift of 256 samples. A short-time Fourier analysis is used to compute the DFT of each overlapping windowed frame. Then the 257-dimensional LPS features normalized by global mean and variance are used to train neural networks. 1024 cells are used for each LSTM layer. The Microsoft Computational Network Toolkit (CNTK) [34] is used for training. For progressive learning systems, one LSTM layer is used to connect the input layer and target layers. According to different SNR gaps, the number of learning targets in progressive learning can be set to 2, 3, 5 and 7. The detailed target SNR gain configurations are shown in Table 1. The

Table 1. Target SNR gain configurations for the intermediate targets in progressive learning systems.

System	SNR Gains for the intermediate targets
2 Targets	10dB (Target 1)
3 Targets	10dB (Target 1-2)
5 Targets	5dB (Target 1-4)
7 Targets	2.5dB (Target 1-4), 5dB (Target 5-6)

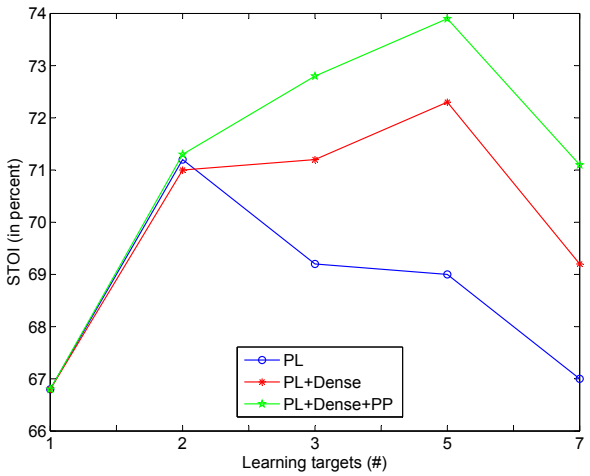

Fig. 4. The average STOI performances of LSTM-based progressive learning systems along with different learning targets across six unseen testing noises at -5dB.

parameter α_k in Eq. 6 is set as follows: $\alpha_K = 1.0; \alpha_k = 0.1, (k = 1, ..., K-1)$.

The motivation of progressive learning is to improve the speech intelligibility in low SNR environments. Fig. 4 shows the STOI performances of LSTM-based progressive learning (PL), densely connected progressive learning (PL+Dense) and post-processed PL+Dense (PL+Dense+PP) along with different learning targets across six unseen testing noises at -5dB. It should be noted that, when the number of learning targets is 1, the result actually corresponds to LSTM baseline system. In Fig. 4, several observations could be made. First, we focus on the blue line, which represents the STOI performance of PL along with the number of learning targets increasing. We can observe that PL achieved a significant STOI improvement from LSTM baseline system to PL system with two learning targets. However, more learning targets could not yield performance gains and instead lead to performance degradation. This might be explained as the information loss in deeper target layer because the dimension of each target layer (257) is much smaller than that of each LSTM layer (1024). In reaction to the phenomenon, we modify conventional PL to densely connected PL. The red line in Fig. 4 shows the results of PL+Dense. Densely connected PL can make full use of the rich set of information from the multiple learning targets, yielding performance improvements in deeper network with more learning targets. PL+Dense obtained the best result when 5 targets are learned. 7 learning targets caused a sharp STOI degradation because there are too many layers (15 layers) in this case and the dimension of the concatenated vector of multiple targets in the dense structure is quite high (1799 dimension), which is with the risk of overfitting. Finally, we focus on the green line which represents PL+Dense+PP. More learning targets produced more information

Table 2. The average STOI and SDR comparison of different systems across six unseen noises at -5dB, 0dB, 5dB and 10dB. The corresponding model size (N_M in MB) and the number of hidden layers (N_H) are also presented.

			STOI (in percent)			
System	N_H	N_M	-5dB	0dB	5dB	10dB
Noisy	-	-	64.7	76.5	86.7	93.2
LSTM Baseline	2	53.0	66.8	80.2	88.4	92.9
	3	85.0	67.3	81.1	89.0	93.1
	4	117.0	67.8	81.1	88.9	93.0
PL	9	105.0	69.0	82.9	90.2	93.0
PL+Dense	9	145.0	72.3	84.5	91.5	95.1
PL+Dense+PP	9	145.0	**73.9**	**85.1**	**91.9**	**95.7**
			SDR			
System	N_H	N_M	-5dB	0dB	5dB	10dB
Noisy	-	-	-6.26	-1.32	3.66	8.65
LSTM Baseline	2	53.0	2.01	5.76	8.59	10.66
	3	85.0	2.32	6.05	8.81	10.80
	4	117.0	2.31	6.00	8.71	10.67
PL	9	105.0	2.69	6.56	9.24	10.64
PL+Dense	9	145.0	3.24	7.42	10.66	13.24
PL+Dense+PP	9	145.0	**3.52**	**7.95**	**11.57**	**14.81**

available for post-processing. When the number of learning targets is larger than 2, PP can further improve the speech intelligibility. In the following experiments, 5 learning targets will be applied for progressive learning.

Table 2 lists the average STOI and SDR results of different systems across six unseen noise types at -5dB, 0dB, 5dB and 10dB. The corresponding model size and the number of hidden layers are also presented. We first focus on LSTM Baseline systems which have been implemented with different hidden layers. With the increase of the number of hidden layer, the performance of LSTM Baseline was easily saturated for both STOI and SDR. When PL was implemented, better results were obtained by a deeper network with 9 hidden layers, e.g., STOI from 67.8 to 69.0 and SDR from 2.31 to 2.69 at -5dB, compared with LSTM Baseline with 4 hidden layers (similar model size to PL). With the dense structure, quite significant gains could be achieved (PL vs. PL+Dense), especially for STOI gain at low SNR (3.3 at -5dB input) and SDR gain at high SNR (2.6dB at 10dB input). Furthermore, PL+Dense was still quite effective to generate 1.9 STOI gain over the unprocessed system (Noisy) by considering that the other systems (LSTM Baseline and PL) failed to improve STOI at 10dB case. Finally, the post-processing can generate additionally remarkable improvements for all SNR levels.

Fig. 5 shows spectrograms of an utterance corrupted by factory noise at -5dB SNR and enhanced by LSTM Baseline, PL and PL+Dense. The LSTM Baseline can achieve a good noise reduction but with severe speech distortion and speech loss. Meanwhile, PL could generate the enhanced speech with less speech distortion, for example, as shown in Fig. 5 (d), the yellow dotted box area. The more severe speech loss problem has been alleviated by PL+Dense compared with PL and LSTM Baseline, as shown in Fig. 5 (e), the yellow dotted box area, demonstrating the effectiveness of dense structure.

5. CONCLUSION

In this study, we explore densely connected progressive learning for LSTM-based speech enhancement to improve the speech intelligibility. The direct mapping from noisy to clean speech is decomposed into multiple stages with SNR increasing progressively by guiding

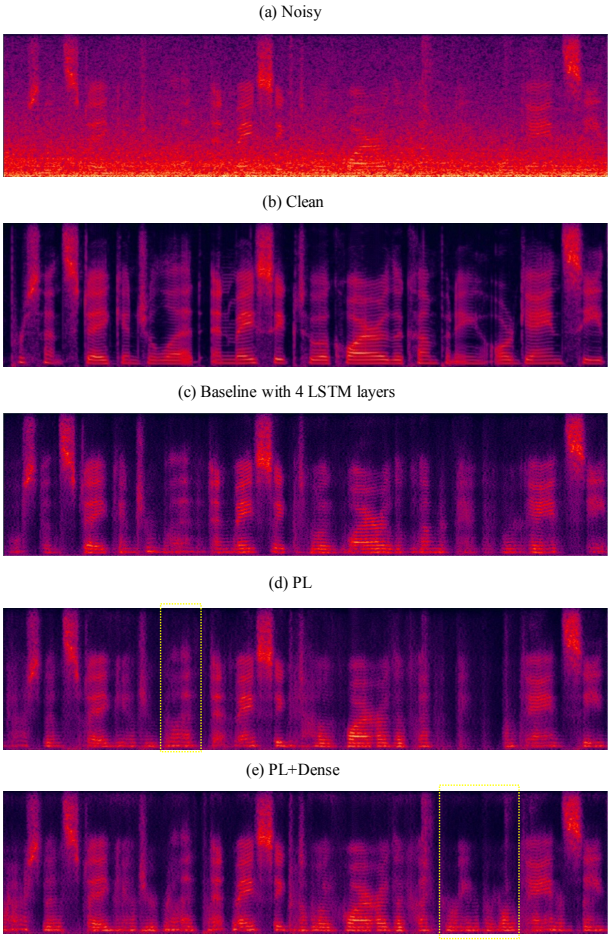

Fig. 5. Spectrograms of an utterance corrupted by factory noise at -5dB SNR: (a) Noisy speech, (b) Clean speech, (c) LSTM Baseline with 4 hidden layers, (d) PL, (e) PL+Dense.

hidden layers in the LSTM network to learn targets explicitly. In order to make full use of the rich set of information from the multiple learning targets, we propose densely connected progressive learning in which the input and the estimations of intermediate target are spliced together to learn next target. Experimental results demonstrate that densely connected progressive learning can learn more targets and yield significantly better speech intelligibility for both low and high SNRs. Moreover, the proposed enhancement approach also demonstrate it effectiveness as a preprocessor in the speaker diarization task under adverse acoustic conditions [35]. In future, the detailed analysis of the learning process in densely connected progressive learning and more testing of the generalization capability will be explored.

6. ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005, and MOE-Microsoft Key Laboratory of USTC. This work was also supported by Samsung.

7. REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[3] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, 1985.

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

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

[7] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments," *Speech Communication*, vol. 95, pp. 28–39, 2017.

[8] P. Mowlaee and J. Kulmer, "Phase estimation in single-channel speech enhancement: Limits-potential," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 8, pp. 1283–1294, 2015.

[9] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.

[10] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 2140–2151, 2013.

[11] H. T. Fan, J. Hung, X. Lu, S. S. Wang, and Y. Tsao, "Speech enhancement using segmental nonnegative matrix factorization," in *ICASSP*, 2014, pp. 4483–4487.

[12] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *Signal Processing Letters, IEEE*, vol. 21, no. 1, pp. 65–68, 2014.

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

[14] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *INTERSPEECH*, 2008, pp. 569–572.

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

[16] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 12, pp. 2136–2147, 2015.

[17] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," 1999.

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

[19] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[20] S. W. Fu, Y. Tsao, and X. Lu, "SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement," in *INTERSPEECH*, 2016, pp. 3768–3772.

[21] S. W. Fu, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *arXiv preprint arXiv:1709.03658*, 2017.

[22] T. Gao, J. Du, Y. Xu, C. Liu, L.-R. Dai, and C.-H. Lee, "Improving deep neural network based speech enhancement in low SNR environments," in *Latent Variable Analysis and Signal Separation*, pp. 75–82. Springer, 2015.

[23] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in *INTERSPEECH*, 2015, pp. 1508–1512.

[24] W. Jiang, H. Zheng, S. Nie, and W. Liu, "Multiscale collaborative speech denoising based on deep stacking network," in *IJCNN*, 2015, pp. 1–5.

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

[26] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Snr-based progressive learning of deep neural network for speech enhancement.," in *INTERSPEECH*, 2016, pp. 3713–3717.

[27] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*. IEEE, 2013, pp. 6645–6649.

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

[29] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, 2013, pp. 1310–1318.

[30] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[31] A. Varga and H. J. 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 Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[32] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *ICASSP*, 2010, pp. 4214–4217.

[33] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[34] 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.

[35] S. Lei, D. Jun, et al., "A novel LSTM-based speech preprocessor for speaker diarization in realistic mismatch conditions," in *Accepted by ICASSP*, 2018.