INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

SNR-Based Progressive Learning of Deep Neural Network for 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

In this paper, we propose a novel progressive learning (PL) framework for deep neural network (DNN) based speech enhancement. It aims at decomposing the complicated regression problem of mapping noisy to clean speech into a series of subproblems for enhancing system performances and reducing model complexities. As an illustration, we design a signal-to-noise ratio (SNR) based PL architecture by guiding each hidden layer of the DNN to learn an intermediate target with gradual SNR gains explicitly. Furthermore, post-processing, with the rich set of information from the multiple learning targets, can further be conducted. Experimental results demonstrate that SNR-based progressive learning can effectively improve perceptual evaluation of speech quality and short-time objective intelligibility in low SNR environments, and reduce the model parameters by 50% when compared with the DNN baseline system. Moreover, when combined with post-processing, the proposed approach can be further improved.

Index Terms: speech enhancement, SNR, progressive learning, deep neural networks, nonlinear regression

1. Introduction

Single channel speech enhancement has been an open research problem for a long time. The goal of speech enhancement is to improve the speech quality and intelligibility in the presence of an interfering noise signal [1]. The background noise can cause performance degradation for real-world applications, including speech communication, hearing aids and speech recognition [2]. Many algorithms have been proposed to solve this problem, and they can be classified into two categories, namely unsupervised and supervised methods.

As for unsupervised approaches, there are, spectral subtraction [3], Wiener filtering [4, 5], minimum mean squared error (MMSE) estimation [6] and optimally-modified log-spectral amplitude (OM-LSA) speech estimator [7, 8]. However, many assumptions were made during the derivation process of these solutions. The noise tracking capacity is limited for highly nonstationary noise cases, and the resulting enhanced speech often suffers from an annoying artifact called musical noise. More recently, some phase-aware speech enhancement methods were investigated in [9, 10, 11].

Supervised and unsupervised nonnegative matrix factorization (NMF) methods were investigated in [12, 13] for speech enhancement. The basic idea is to decompose the noisy speech data into bases and weights matrices for the speech and noise, respectively. On the other hand, supervised deep learning approaches have also been developed in recent years. The applications of DNN in speech signal processing area, create a new direction of single channel speech enhancement. In [14, 15], masking techniques were used to make binary classification on time-frequency (T-F) units for speech separation. Xu *et, al.* [16, 17] proposed a DNN-based speech enhancement framework in which DNN was regarded as a regression model to predict the clean log-power spectra (LPS) features [18] from noisy LPS features. In [19, 20], DNN-based method was demonstrated to be more effective than the NMF-based method in speech separation. In [21], we proposed a joint framework combining speech enhancement with voice activity detection (VAD) to increase the speech intelligibility in low SNR environments. In [22], Long Short-Term Memory (LSTM) based speech enhancement was explored. In [23], Kim *et, al.* aimed at a fine-tuning scheme at the test stage to improve the performance of a well-trained Denoising AutoEncoder (DAE).

From the view of machine learning, the challenge of DNN-based speech enhancement is the optimization of the complicated and non-convex objective function. Recently, multi-task learning (MTL) [24] has been adopted in speech enhancement. In [25], a multi-objective framework was proposed to improve the generalization capability of regression DNN. Based on MTL method, Jiang *et, al.* [26] proposed a framework to improve 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 [27] originated from cognitive science. The basic idea is to start small, learn easier aspects of the task or easier sub-tasks, and then gradually increase the difficulty level. Curriculum learning is related with 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.

In this paper, based on previous work and inspired by curriculum learning, we propose a progressive learning framework to improve the performance of DNN-based speech enhancement especially in low SNR environments. As a demonstration of DNN training, the direct mapping from the noisy speech to clean speech is decomposed into multiple stages with SNR increasing progressively. We guide hidden layers to learn targets explicitly, which can significantly reduce the model complexity. And the subproblem solving in each stage can boost the subsequent learning of the next stage. Furthermore, the estimated targets of different stages provide rich information for multi-target fusion as a post-processing. Experimental results demonstrate that the proposed approach can not only significantly improve both objective measures of speech quality and intelligibility but

http://dx.doi.org/10.21437/Interspeech.2016-224

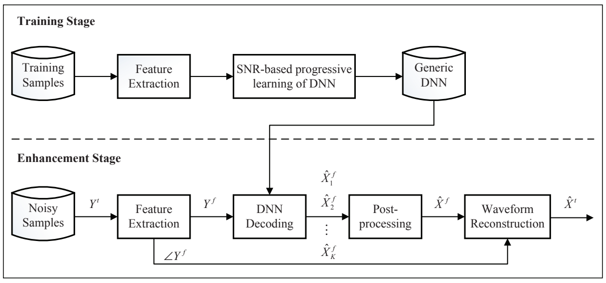


Figure 1: *Overall development flow and architecture.*

also reduce model parameters by 50% when compared with the conventional DNN.

2. System Overview

The overall flowchart of our proposed SNR-based progressive learning framework for speech enhancement is illustrated in Figure 1. In the training stage, a regression DNN model is progressively trained from a collection of stereo data, consisting of pairs of noisy speech at different levels of SNR and clean speech represented by LPS features. In the enhancement stage, the well-trained DNN model is fed with the noisy features in order to generate multiple enhanced LPS features ($\hat{X}_1^{\text{f}}, \hat{X}_2^{\text{f}} ... \hat{X}_K^{\text{f}}$) of different SNR levels. Another module, namely post-processing, is proposed to perform the fusion of the multiple estimations. The additional phase information is calculated from the original noisy speech. Finally an overlap-add method is used to synthesize the waveform of the enhanced speech. A detailed description of the feature extraction module and waveform reconstruction module can be found in [18].

3. SNR-based Progressive Learning

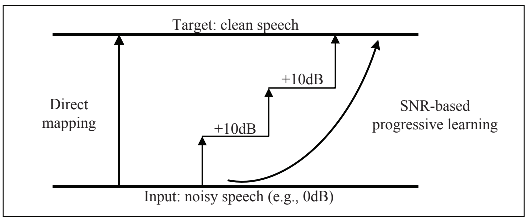

Target: clean speech
Direct mapping
+10dB
+10dB
SNR-based progressive learning
Input: noisy speech (e.g., 0dB)

Figure 2: *Illustration of SNR-based progressive learning.*

3.1. Motivation

Although DNN has been successfully adopted as a regression model for speech enhancement, the resulting enhanced speech often suffers from speech distortion in low SNR environments. On the other hand, the conventional microphone array aims to achieve the SNR gain of input noisy speech with less speech distortion, which should be complementary to the direct learning of the clean speech as the targets with potentially more speech distortion in DNN-based speech enhancement, especially in low SNR environments. Further inspired by curriculum learning, SNR-based progressive learning is proposed, as shown in Figure 2. 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. For example, the input SNR of noisy speech is 0dB, then two intermediate learning targets are 10dB and 20dB speech while the final target is the clean speech (infinity dB).

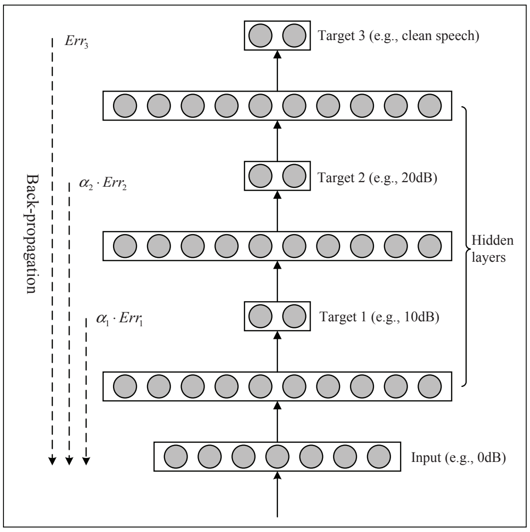


Figure 3: *DNN architecture for speech enhancement.*

3.2. DNN Training

In [16], DNN acted as a regression model to predict the clean LPS features given the input noisy LPS features with acoustic context. In this study, the DNN architecture for SNR-based progressive learning is illustrated in Figure 3. The active function is linear in the target layers and sigmoidal in the other hidden layers. All the target layers are designed to learn intermediate speech with higher SNRs or clean speech. This stacking style DNN can learn multiple targets progressively and efficiently. In the forward pass, the enhanced features of the current target layer are used as the input of the next target layer. Then, the back-propagation algorithm is adopted with the MMSE criteria defined for the K ($K = 3$) target layers (Err_1, Err_2, Err_3) with the same form of objective function as follows,

$$Err = \frac{1}{N}\sum_{n=1}^{N}(\|\hat{X}_n^t - X_n^t\|_2^2) \tag{1}$$

where $\hat{X}_n^t$ and X_n^t are the n^{th} D-dimensional vectors of estimated and reference target features, respectively, with N representing the mini-batch size. Err_1, Err_2 and Err_3 will be combined together to calculate the back-propagated gradients in a weighted sum fashion as:

$$\boldsymbol{\epsilon} = \frac{\partial(Err_3)}{\underset{1\le\ell\le L_3+1}{\partial(\boldsymbol{W}^\ell, \boldsymbol{b}^\ell)}} + \alpha_2\frac{\partial(Err_2)}{\underset{1\le\ell\le L_2+1}{\partial(\boldsymbol{W}^\ell, \boldsymbol{b}^\ell)}} + \alpha_1\frac{\partial(Err_1)}{\underset{1\le\ell\le L_1+1}{\partial(\boldsymbol{W}^\ell, \boldsymbol{b}^\ell)}} \tag{2}$$

where $\boldsymbol{\epsilon}$ is the overall gradient of the objective function with $(\boldsymbol{W}^\ell, \boldsymbol{b}^\ell)$ denoting the weights and bias parameters to be learned at the ℓ-th layer, L_1, L_2 and L_3 representing the number of hidden layers between the input layer and each target layer. The gradients from each target layer only affect the parameters update of its front-end layers. α_1 and α_2 are weighting factors to balance multiple targets. If $\alpha_1 = \alpha_2 = 0$, it is similar to the conventional DNN with a low-rank structure [28]. In this paper, we set $\alpha_1 = \alpha_2 = 0.1$.

3.3. Post-processing

An important benefit from SNR-based progressive learning is the estimated features (Out_1, Out_2, Out_3) of different targets provide rich information for post-processing. Out_1, Out_2 and Out_3 make different tradeoffs between more noise reduction and less speech distortion in different input SNR conditions. In this study, we simply average these estimated features to further improve the overall performance.

4. Experiments and Result Analysis

First, 115 noise types used in [25] were chosen as our noise database, including 100 noise types [29] and home-made musical noises. Clean speech is derived from the WSJ0 corpus [30]. 7138 utterances (about 12 hours of read speech) from 83 speakers, denoted as SI-84 training set, were corrupted with the above mentioned 115 noise types at different SNR levels, i.e., -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, were used to construct the test set for each combination of noise types and SNR levels (-5dB, 0dB, 5dB). Three unseen noises from the NOISEX-92 corpus [31], namely Babble, Factory and Destroyer engine, were adopted for testing.

As for signal analysis, 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 were used to train DNNs. PESQ [32] and STOI [33] were used to assess the quality and intelligibility of the enhanced speech. For DNN training, global mean and variance normalization was applied to the input and output reference feature vectors, and the DNN was initialized with random weights. A configuration with 3 hidden layers, 2048 sigmoidal units at each hidden layer, 7-frame input and 1-frame output was used to train our DNN baseline system. The DNN architecture for SNR-based progressive learning was 1799-2048-257-2048-257-2048-257, denoting 7-frame input and 1-frame output in target layers. According to the SNR diversity of the input data, two sets of experiments, namely single-SNR and multi-SNR training were designed as follows.

4.1. Single-SNR training

For single-SNR training part, the input data contains only one SNR level. Table 1 lists the SNR configuration of single-SNR training for progressive learning. For example, if the input speech was at -5dB SNR, the three learning targets were set as 5dB, 15dB and clean speech. And for DNN baseline system, the learning target was clean speech. Table 2 gives a detailed PESQ and STOI comparison of different systems on the test set at 0dB of three unseen noise environments. ”Noisy” and ”Baseline DNN (12.6M)” represent the systems of original noisy speech and the conventional DNN for speech enhancement with 12.6 million weight parameters, respectively. ”SNR-PL DNN: Out1”, ”SNR-PL DNN: Out2” and ”SNR-PL DNN: Out3” are estimations of noisy speech at 10dB, 20dB and clean speech. ”SNR-PL DNN: PP (6.3M)” denotes SNR-based progressive learning combined with post-processing.

Table 1: Target SNR configurations of progressive learning for single-SNR training.

Input	Target 1	Target 2	Target 3
-5dB	5dB	15dB	clean speech
0dB	10dB	20dB	clean speech
5dB	15dB	25dB	clean speech

From Table 2, several observations could be made. First, the baseline DNN system could improve PESQ consistently over the unprocessed system while STOI was degraded across three noise types, which implied that the baseline DNN introduced some perceptible speech distortions at low SNRs. However, the intermediate results of SNR-based progressive learning provided rich information for the analysis in comparison to the conventional DNN training. At the first stage of SNR-based progressive learning, Out_1 could improve both PESQ and STOI compared with the noisy speech results, which indicated that the direct mapping from noisy speech at low SNR to clean speech might not be satisfactory in real practice due to its complicated relationship to be learned. Then, Out_2 achieved additional gains over Out_1 in most cases. As for the final stage, the performance of Out_3 was degraded when compared with Out_2 due to a large span of SNR increase, but Out_3 still outperformed DNN baseline in terms of both speech quality and intelligibility. Based on simply average operation as the post-processing, our final result SNR-PL DNN: PP could take advantage of Out_1, Out_2 and Out_3 to further improve the overall performance. Compared with the results of DNN baseline, SNR-PL DNN: PP not only yielded significant improvements of PESQ and STOI across all noise types but also reduced parameters by 50%.

Table 3 also lists the results of different single-SNR training systems for -5dB and 5dB on the test set of three unseen noise environments. In comparison to the 0dB case in Table 2, our proposed approach was still quite effective for all measures at the lower SNR while remarkable gains could be achieved especially in STOI measure at the higher SNR.

4.2. Multi-SNR training

In [16, 17, 25], all experiments were conducted in multi-SNR training style with the input noisy speech at different SNR levels. For a fair comparison to further demonstrate the effectiveness of the progressively trained DNNs, we also design the experiments for multi-SNR training conditions in the following. The input and target features at different SNRs in Table 1 for every learning stage were combined for DNN training. The first and second stages of progressive learning aimed at generating a 10dB SNR gain for the input speech with different SNRs. Table 4 shows the results for multi-SNR training on the test set at -5dB and 0dB SNR. Obviously, the performance of the baseline was not satisfactory in low SNR environments. However, the performance of SNR-PL DNN was consistent with that in single-SNR training, i.e., significantly outperforming noisy speech and the DNN baseline, especially for speech intelligibility.

Figure 4 shows spectrograms of an utterance corrupted by Destroyer engine noise at -5dB SNR and enhanced by multi-SNR training systems. The conventional DNN can achieve a good noise reduction but with severe speech distortion. Meanwhile, our proposed approach could generate the enhanced speech with less speech distortion, for example, as shown in the three solid line box areas. Furthermore, although post-processing retained more background noises, speech distortion

Table 2: A detailed PESQ and STOI comparison of different single-SNR training systems at 0dB SNR on the test set of three unseen noise environments (N1: Babble, N2: Factory, N3: Destroyer engine), among: Noisy, DNN baseline, estimations of different levels of SNR and SNR-based progressive learning combined with post-processing (denoted as SNR-PL DNN: PP).

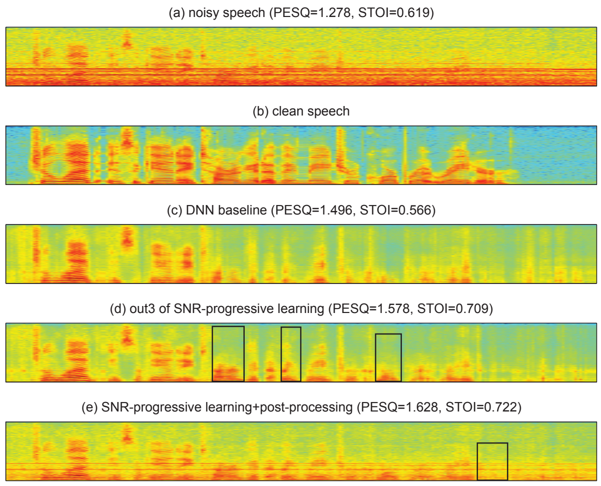
Single-SNR training						
	N1 (0dB)		N2 (0dB)		N3 (0dB)	
System	PESQ	STOI	PESQ	STOI	PESQ	STOI
Noisy	1.683	0.711	1.689	0.757	1.636	0.749
Baseline DNN (12.6M)	1.775	0.710	1.875	0.702	1.760	0.694
SNR-PL DNN: Out1	1.828	0.730	1.850	0.764	1.693	0.763
SNR-PL DNN: Out2	2.015	0.747	2.023	0.764	1.866	0.757
SNR-PL DNN: Out3	1.789	0.731	1.894	0.722	1.760	0.710
SNR-PL DNN: PP (6.3M)	2.007	0.766	2.017	0.783	1.928	0.781

Table 3: PESQ and STOI comparison of different single-SNR training systems for -5dB and 5dB cases on the test set of three unseen noise environments (N1: Babble, N2: Factory, N3: Destroyer engine).

Single-SNR training												
	N1 (-5dB)		N2 (-5dB)		N3 (-5dB)		N1 (5dB)		N2 (5dB)		N3 (5dB)	
System	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
Noisy	1.449	0.587	1.387	0.634	1.422	0.627	2.002	0.824	2.032	0.862	1.899	0.853
Baseline DNN (12.6M)	1.156	0.531	1.468	0.562	1.247	0.523	2.367	0.834	2.391	0.825	2.323	0.824
SNR-PL DNN: PP (6.3M)	1.514	0.618	1.550	0.648	1.414	0.637	2.369	0.864	2.431	0.878	2.352	0.879

Table 4: PESQ and STOI comparison for multi-SNR training system at -5dB and 0dB SNR on the test set of three unseen noise environments (N1: Babble, N2: Factory, N3: Destroyer engine).

Multi-SNR training												
	N1 (-5dB)		N2 (-5dB)		N3 (-5dB)		N1 (0dB)		N2 (0dB)		N3 (0dB)	
System	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
Noisy	1.449	0.587	1.387	0.634	1.422	0.627	1.683	0.711	1.689	0.757	1.636	0.749
Baseline DNN (12.6M)	1.371	0.582	1.599	0.625	1.396	0.583	1.961	0.742	2.090	0.761	1.924	0.732
SNR-PL DNN: PP (6.3M)	1.545	0.630	1.690	0.683	1.541	0.673	2.053	0.771	2.147	0.800	1.999	0.797


(a) noisy speech (PESQ=1.278, STOI=0.619)

(b) clean speech

(c) DNN baseline (PESQ=1.496, STOI=0.566)

(d) out3 of SNR-progressive learning (PESQ=1.578, STOI=0.709)

(e) SNR-progressive learning+post-processing (PESQ=1.628, STOI=0.722)

Figure 4: *Spectrograms of an utterance corrupted by Destroyer engine noise at -5dB SNR and enhanced by multi-SNR training: (a) noisy speech, (b) clean speech, (c) DNN baseline (PESQ=1.496, STOI=0.566); (d) out3 in the proposed DNN (PESQ=1.578, STOI=0.709); (e) further post-processing (PESQ=1.628, STOI=0.722).*

could be further reduced especially in the speech segment (box area in Figure 4 (e)) with quite low SNR, which improved both speech quality (PESQ) and speech intelligibility (STOI).

5. Conclusions

In this study, we propose a novel SNR-based progressive learning framework to improve the performance of regression DNN based speech enhancement in low SNR environments. The direct mapping from noisy to clean speech is decomposed into multiple stages with SNR increasing progressively by guiding hidden layers in the DNN architecture to learn targets explicitly. We test the effectiveness of the proposed framework in single-SNR and multi-SNR training conditions under three unseen noise environments. Experimental results demonstrate that this approach can effectively improve the enhancement performance and reduce parameters by 50% when compared with the conventional DNN approach. Furthermore, multiple estimated targets provide rich information for post-processing. The simple average operation as post-processing can further generate significant performance gains, especially for speech intelligibility. In the future, other progressive learning strategies combined with post-processing will be further explored.

6. Acknowledgements

This work was partially funded by the National Natural Science Foundation of China under Grants No. 61305002 and MOE-Microsoft Key Laboratory of USTC.

7. References

[1] J. Benesty, S. Makino, and J. D. Chen, *Speech enhancement*. Springer, 2005.

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

[3] 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.

[4] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 3, pp. 197–210, 1978.

[5] ——, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[6] 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.

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

[8] 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.

[9] 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.

[10] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: history and recent advances," *Signal Processing Magazine, IEEE*, vol. 32, no. 2, pp. 55–66, 2015.

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

[12] 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.

[13] 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.

[14] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *ICASSP*, 2013, pp. 7092–7096.

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

[16] 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.

[17] ——, "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.

[18] 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.

[19] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *ICASSP*, 2014, pp. 1562–1566.

[20] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *INTERSPEECH*, 2014, pp. 2685–2689.

[21] 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*. Springer, 2015, pp. 75–82.

[22] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, L. 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*. Springer, 2015, pp. 91–99.

[23] M. Kim and P. Smaragdis, "Adaptive denoising autoencoders: A fine-tuning scheme to learn from test mixtures," in *Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 100–107.

[24] R. Camana, "Multitask learning: A knowledge-based source of inductive bias," in *ICML*, 1993, pp. 41–48.

[25] 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.

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

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

[28] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in *ICASSP*, 2013, pp. 6655–6659.

[29] G. Hu, "100 nonspeech environmental sounds," 2004.

[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] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001, pp. 749–752.

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