

GAUSSIAN DENSITY GUIDED DEEP NEURAL NETWORK FOR SINGLE-CHANNEL SPEECH ENHANCEMENT

Li Chai, Jun Du, Yan-nan Wang

University of Science and Technology of China, Hefei, Anhui, P. R. China
cl122@mail.ustc.edu.cn, jundu@ustc.edu.cn, wyn314@mail.ustc.edu.cn

ABSTRACT

Recently, the minimum mean squared error (MMSE) has been a benchmark of optimization criterion for deep neural network (DNN) based speech enhancement. In this study, a probabilistic learning framework to estimate the DNN parameters for single-channel speech enhancement is proposed. First, the statistical analysis shows that the prediction error vector at the DNN output well follows a unimodal density for each log-power spectral component. Accordingly, we present a maximum likelihood (ML) approach to DNN parameter learning by characterizing the prediction error vector as a multivariate Gaussian density with a zero mean vector and an unknown covariance matrix. It is demonstrated that the proposed learning approach can achieve a better generalization capability than MMSE-based DNN learning for unseen noise types, which can significantly reduce the speech distortions in low SNR environments.

Index Terms— Prediction error modeling, multivariate Gaussian density, maximum likelihood estimation, deep neural network, speech enhancement

1. INTRODUCTION

Speech enhancement [1] is an important front-end of speech processing systems aiming at noise reduction. The background noise can reduce both the quality and intelligibility of the speech signals, and cause performance degradations for real-world applications, such as automatic speech recognition (ASR), mobile communication and hearing aids. The problem of enhancing noisy speech recorded by a single microphone has attracted a considerable amount of research attention [2]. Considering the process of noise corruption on speech is very complicated, the performance of speech enhancement in real acoustic environments is still unsatisfactory and many issues should be explored.

Numerous speech enhancement methods were developed over the past several decades. The conventional methods include a wide range of approaches, such as spectral subtraction [3], Wiener filtering [4, 5], a MMSE estimator [6], an optimally-modified log-spectral amplitude (OM-LSA) speech

estimator [7, 8] and so on. A common problem usually encountered in these conventional methods is that the resulting enhanced speech often suffers from an annoying artifact called musical noise [9]. In addition, they often fail to track non-stationary noise for real-world scenarios in unexpected acoustic conditions.

Following the successes in deep learning based speech research, such as speech recognition, speech separation and speech enhancement [10, 11, 12], Xu et al. proposed a regression deep neural network (DNN) based speech enhancement framework [13, 14] which was adopted to model the complicated relationship between the noisy speech and clean speech via training a deep and wide neural network architecture using a large collection of heterogeneous training data and the abundant acoustic context information. In contrast with conventional approaches, it makes no assumptions about the statistical properties of the signals. In addition, it can also handle non-linear and highly non-stationary noises effectively. However, one problem of this approach we should pay attention to is that some distortions are introduced to the estimated clean speech signal especially in low SNR environments because the regression DNN removes the noise considerably from the noisy speech. To address this issue, some work has been done, as shown in [15, 16, 17]. More complex neural network structure for speech enhancement has great attractions for many researchers. Sun proposed a separable deep auto encoder (SDAE) to estimate the spectrum of clean speech and noise respectively by minimizing the total reconstruction error of noisy speech spectrum [18]. In [19], Mass et al. introduced deep recurrent neural networks (DRNNs) as an approach for feature enhancement in robust ASR. In [20], long short-term memory (LSTM) recurrent neural network based speech enhancement was explored. In addition, convolutional neural network (CNN) was investigated in [21].

One challenge of DNN-based speech enhancement is the optimization of the complicated and non-convex objective function from the view of machine learning. The MMSE between the target features and the predicted features was commonly used as the objective function which performed better than many other objective functions [22], such as the Kullback Leibler divergence [23] or the Itakura-Saito diver-

gence [24]. In [25], a multi-objective learning framework was proposed to optimize a joint objective function, encompassing MMSE not only in the primary clean log-power spectra (LPS) features but also in secondary targets for continuous features and categorical information. This joint optimization of different but related targets can potentially improve the DNN prediction performance of the primary target. In this paper, we explore a maximum likelihood (ML) criterion within the probabilistic learning framework to optimize DNN parameters with the assumption that the prediction error vector of the regression DNN follows a multivariate Gaussian density. Accordingly, a training procedure of ML-based DNN (ML-DNN) is designed to update both DNN parameters and the covariance matrix of Gaussian density alternatively. The MMSE-based DNN (MMSE-DNN) approach could be thought as a special case of the proposed ML-DNN approach with an identity covariance matrix. The evaluation on the TIMIT corpus [26] shows that the proposed ML-DNN approach achieves a significantly better improvement of speech quality and intelligibility than the conventional MMSE-DNN approach. Moreover, it has a better generalization capability and achieves less speech distortions.

2. THE PROPOSED ML-DNN APPROACH

We redefine the objective function in the probabilistic framework and adopt the maximum likelihood estimation for the parameter learning aiming at further improving the generalization capability of the conventional MMSE optimization for the regression DNN, as shown in Fig. 1. The input of DNN is the D -dimensional LPS feature vector of noisy speech with an acoustic context of $(2\tau + 1)$ neighboring frames.

In conventional MMSE-DNN, a mini-batch stochastic gradient descent algorithm is performed in mini-batches with multiple epochs to improve the following error function,

$$E = \frac{1}{N} \sum_{n=1}^N \|(\hat{\mathbf{x}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W}) - \mathbf{x}_n)\|_2^2, \quad (1)$$

where E is the mean squared error, $\hat{\mathbf{x}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W})$ and \mathbf{x}_n denote the estimated and reference normalized LPS at sample index n , respectively, with N representing the mini-batch size, $\mathbf{y}_{n-\tau}^{n+\tau}$ being the noisy LPS feature vector where the window size of context is $(2\tau + 1)$, \mathbf{W} denoting the parameters to be learned. The prediction error vector \mathbf{e}_n at sample index n could be defined as:

$$\mathbf{e}_n = \mathbf{x}_n - \hat{\mathbf{x}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W}). \quad (2)$$

We make an assumption that it follows a multivariate Gaussian density with a D -dimensional zero mean vector and a $D \times D$ unrestricted covariance matrix Σ :

$$p(\mathbf{e}_n) = \mathcal{N}(\mathbf{e}_n | \mathbf{0}, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \mathbf{e}_n^\top \Sigma^{-1} \mathbf{e}_n\right). \quad (3)$$

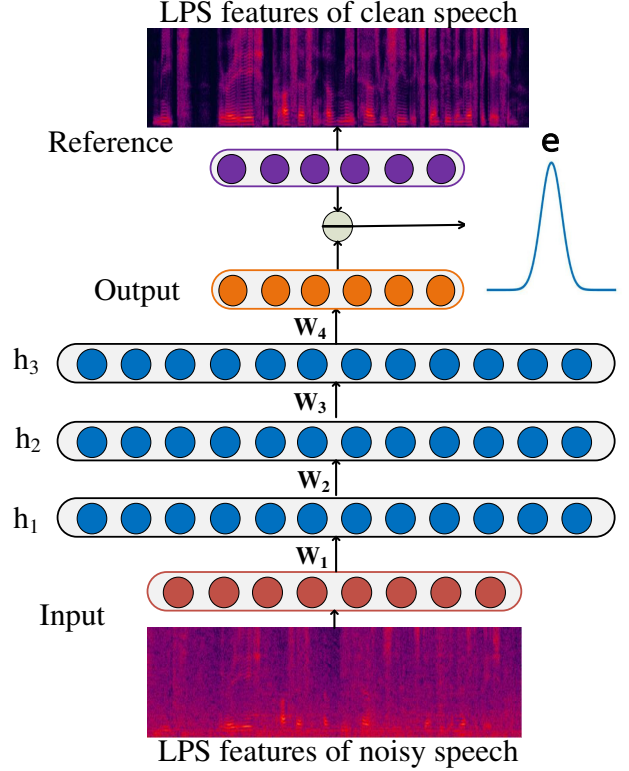


Fig. 1. The ML-DNN architecture for speech enhancement.

If the reference vector is also a random vector, then we can get the equivalent expression as Eq. (3):

$$p(\mathbf{x}_n | \mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W}, \Sigma) = \mathcal{N}(\mathbf{x}_n | \hat{\mathbf{x}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W}), \Sigma), \quad (4)$$

which implies that the conditional distribution of \mathbf{x}_n given $\mathbf{y}_{n-\tau}^{n+\tau}$ with the parameter set (\mathbf{W}, Σ) is unimodal. Given a mini-batch set with N data pairs $(\mathbf{Y}, \mathbf{X}) = \{(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{x}_n) | n = 1, 2, \dots, N\}$ and making the assumption that these data pairs are drawn independently from the distribution in Eq. (4), we can define the likelihood function as:

$$p(\mathbf{X} | \mathbf{Y}, \mathbf{W}, \Sigma) = \prod_{n=1}^N \mathcal{N}(\mathbf{x}_n | \hat{\mathbf{x}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W}), \Sigma), \quad (5)$$

where the parameter set (\mathbf{W}, Σ) is to be optimized. Accordingly, the log-likelihood function can be written as:

$$\begin{aligned} \ln p(\mathbf{X} | \mathbf{Y}, \mathbf{W}, \Sigma) &= \sum_{n=1}^N \ln \mathcal{N}(\mathbf{x}_n | \hat{\mathbf{x}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W}), \Sigma) \\ &= C - \frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \hat{\mathbf{x}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W}))^\top \Sigma^{-1} (\mathbf{x}_n - \hat{\mathbf{x}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W})), \end{aligned} \quad (6)$$

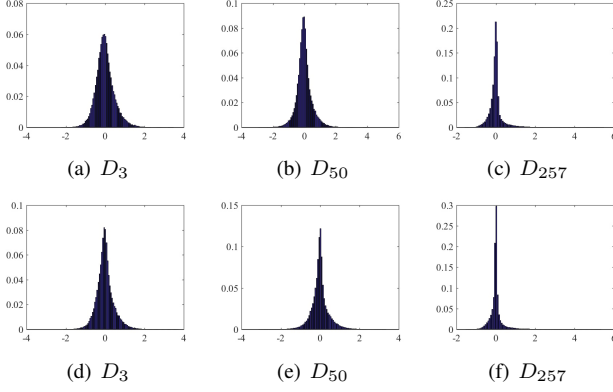


Fig. 2. The distributions for selected dimensions of the prediction error vector from well-trained DNN on the cross validation set: (a)-(c) refer to MMSE-DNN while (d)-(f) correspond to ML-DNN.

where C is a constant. We adopt maximum likelihood criterion to alternatively optimize \mathbf{W} and Σ . To maximize Eq. (6) with respect to \mathbf{W} , it is equivalent to minimizing the following sum-of-squares error function in terms of Mahalanobis distance:

$$M = \sum_{n=1}^N (\mathbf{x}_n - \hat{\mathbf{x}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W}))^\top \Sigma^{-1} (\mathbf{x}_n - \hat{\mathbf{x}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W})). \quad (7)$$

Then the back-propagation procedure with a stochastic gradient descent method is used to optimize \mathbf{W} in the mini-batch mode of N sample frames.

Alternatively, we can also maximize Eq. (6) with respect to Σ . Then the update formula can be derived as:

$$\Sigma = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \hat{\mathbf{x}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W})) (\mathbf{x}_n - \hat{\mathbf{x}}_n(\mathbf{y}_{n-\tau}^{n+\tau}, \mathbf{W}))^\top. \quad (8)$$

Considering the number of free parameters in the distribution, the symmetric covariance matrix Σ has $\frac{D \times (D+1)}{2}$ independent parameters. For large D values, the total number of parameters increases with D in a square form, therefore it is very difficult to calculate the inverse of the covariance matrix Σ . To avoid the problem, we use the diagonal covariance matrix in this study. The whole training procedure is summarized as Algorithm 1.

By comparing Eq. (1) and Eq. (7), we should note that the conventional MMSE-DNN is a special case of ML-DNN where the covariance matrix in Eq. (7) is always an identity matrix, namely making a strong assumption that all the LPS components are with the same variances. However statistical analysis on prediction errors indicates that is not reasonable. We present the distributions of selected dimensions (3, 50, 257) of the prediction error vector on the cross validation set for both well-trained MMSE-DNN and ML-DNN as shown

Algorithm 1 Procedure of ML-DNN training

Step 1: Initialization

Initialize the DNN parameter set \mathbf{W} randomly. The covariance matrix Σ is set to an identity matrix.

Step 2: Fix Σ then update \mathbf{W}

By minimizing Eq. (7), the back-propagation procedure with a stochastic gradient descent method is used to update \mathbf{W} in the mini-batch mode of N sample frames.

Step 3: Fix \mathbf{W} then update Σ

Update Σ via Eq. (8).

Step 4: Go to Step 2 for the next epoch

in Fig. 2. It is observed that all selected dimensions of the prediction error vector approximately follow a unimodal distribution with the mean closing to zero for both MMSE-DNN and ML-DNN, which verifies the reasonability of the zero mean hypothesis. However, the variances are quite different, which indicates that the assumption of equivalent variances in MMSE-DNN is unreasonable. Furthermore, it is observed that the variance of each dimension in ML-DNN is smaller than that in MMSE-DNN, demonstrating that ML-DNN could better model the prediction errors.

3. EXPERIMENTS AND RESULTS

In this paper, all experiments were conducted on waveforms with 16kHz. And 115 noise types were adopted for training to improve the robustness to the unseen noise types. These 115 noise types included 100 noise types recorded by G. Hu [27] and 15 home-made noise types. And the clean speech data was derived from the TIMIT corpus. All 4620 utterances from the training set of the TIMIT database were corrupted with the abovementioned 115 noise types at six levels of signal-to-noise-ratios (SNRs), i.e., 20dB, 15dB, 10dB, 5dB, 0 dB and -5dB, to build 80-hour multi-condition training set, consisting of pairs of clean and noisy speech utterance. The 192 utterances from the core test set of TIMIT database were used to construct the test set for each combination of noise types and SNR levels. In this experiment, three unseen noise types, namely Destroyerops, Factory1 and Pink were adopted for testing. All of them were collected from the NOISEX-92 corpus [28].

A short-time Fourier transform was adopted to compute the spectra of each overlapping windowed frame. Then 257 dimensions ($D=257$) LPS features were used to train DNNs. Sigmoid was used as the activation function of DNN. Mean and variance normalization was applied to the input and target feature vectors of the DNN. All DNN configurations were fixed at $h=3$ hidden layers, 2048 units at each hidden layer, and 7-frame ($\tau=3$) input. For the update of DNN parameters in both MMSE-DNN and ML-DNN, the learning rate for the supervised fine-tuning was set to 0.1 for the first 10 epochs and declined at a rate of 90% after every epoch in the next

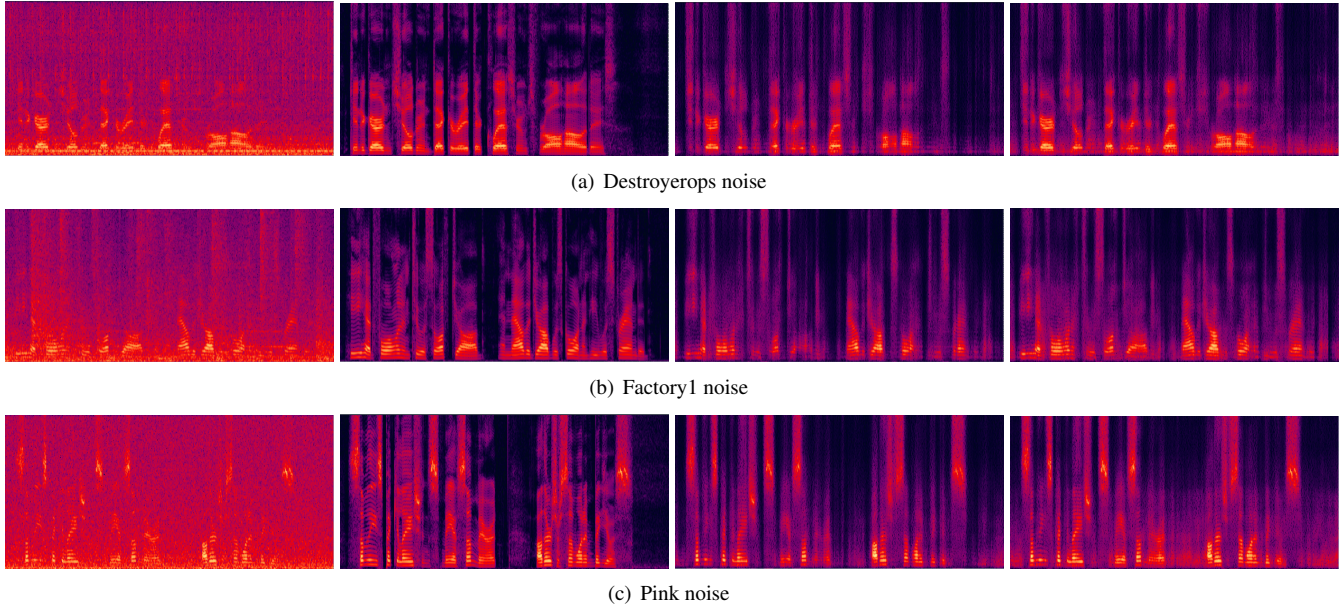


Fig. 3. Comparison of spectrograms of three 16kHz TIMIT utterances corrupted by Destroyerops, Factory1, Pink noise at SNR=0dB respectively (from left to right):noisy speech, clean speech, the enhanced speech by the MMSE-DNN system, the enhanced speech by the ML-DNN system initialized with the well-trained MMSE-DNN

Table 1. Average performance comparison on the test set at different SNRs across three unseen noise environments, among: MMSE-DNN initialized randomly (denoted as MMSE), ML-DNN initialized randomly (denoted as ML1) and ML-DNN initialized with the well-trained MMSE-DNN (denoted as ML2).

SNR (dB)		-5	0	5	10	15	20	Ave
PESQ	MMSE	1.58	2.05	2.48	2.84	3.15	3.43	2.59
	ML1	1.71	2.17	2.59	2.95	3.25	3.53	2.70
	ML2	1.85	2.24	2.61	2.95	3.26	3.54	2.74
STOI	MMSE	0.54	0.68	0.79	0.87	0.92	0.95	0.79
	ML1	0.60	0.73	0.83	0.90	0.94	0.97	0.83
	ML2	0.62	0.74	0.83	0.90	0.94	0.97	0.83
SSNR	MMSE	-1.18	0.20	2.03	4.12	6.23	8.13	3.25
	ML1	-1.44	0.33	2.56	5.05	7.60	10.01	4.02
	ML2	-1.75	0.23	2.63	5.17	7.72	10.08	4.02
LSD	MMSE	6.26	4.60	3.59	2.84	2.32	2.02	3.61
	ML1	5.22	4.14	3.23	2.51	1.93	1.54	3.09
	ML2	5.42	4.18	3.22	2.48	1.92	1.51	3.12

40 epochs with the mini-batch size of 128 ($N=128$). For the waveform reconstruction, the original phase of noisy speech was adopted with the enhanced log-power spectra.

MMSE-DNN was initialized with random weights, while ML-DNN was initialized with random weights and the weights of the well-trained MMSE-DNN respectively. Segmental SNR (SSNR in dB) for measuring noise reduction,

Table 2. Average performance comparison on the test set of three unseen noise environments across different SNRs, among: MMSE-DNN initialized randomly (denoted as MMSE), ML-DNN initialized randomly (denoted as ML1) and ML-DNN initialized with the well-trained MMSE-DNN (denoted as ML2).

		PESQ	STOI	SSNR	LSD
Destroyerops	MMSE	2.58	0.79	3.16	3.80
	ML1	2.73	0.82	3.52	3.15
	ML2	2.74	0.83	3.66	3.11
Factory1	MMSE	2.55	0.79	3.30	3.57
	ML1	2.62	0.82	4.30	3.08
	ML2	2.70	0.83	4.31	3.05
Pink	MMSE	2.63	0.81	3.31	3.45
	ML1	2.75	0.84	4.23	3.06
	ML2	2.79	0.84	4.10	3.20

log-spectral distortion (LSD in dB) for measuring speech distortion [29], perceptual evaluation of speech quality (PESQ) for measuring speech quality [30], and short-time objective intelligibility (STOI) for measuring speech intelligibility [31] are compared in Table 1 and Table 2. Clearly, experimental results demonstrate that the proposed ML-DNN approach yield consistent and significant improvements over the conventional MMSE-DNN approach across the three unseen noise. Finally, the average PESQ, STOI, SSNR and LSD of the ML-DNN system initialized with the well-trained MMSE-

DNN were improved by 0.15, 0.04, 0.8 dB and 0.5 dB respectively in contrast with MMSE-DNN. The improvements of the ML-DNN initialized randomly are smaller than those of the ML-DNN initialized with the well-trained MMSE-DNN. Fig. 3 presented spectrograms of three utterances corrupted by Destroyerops noise, Factory1 noise and Pink noise respectively. It was shown that the ML-DNN approach could yield less speech distortions significantly. Furthermore, please note that larger PESQ and STOI improvements are obtained at low SNRs than those at high SNRs, e.g., ML-DNN system initialized with well-trained MMSE-DNN improved PESQ and STOI by 0.27 and 0.08 respectively over the well-trained MMSE-DNN system at SNR=-5dB, while PESQ and STOI were improved by 0.11 and 0.02 respectively at SNR=20dB. Considering that the MMSE-DNN approach causes speech distortions more or less, especially in low SNR environments, the proposed ML-DNN approach can achieve less speech distortions which is much more significant in low SNR environments.

4. CONCLUSION

In this paper, a novel maximum likelihood approach is proposed to improve DNN training for speech enhancement. The assumption that the prediction error vector of DNN follows the Gaussian distribution was shown to be reasonable. In the ML solution, both the DNN parameters and the covariance matrix of the prediction error vector are jointly and alternatively optimized. Compared with the conventional MMSE optimization, the ML approach could achieve a better generalization capability and less speech distortions.

5. ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 61671422 and U1613211, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDB02070006, and in part by the National Key Research and Development Program of China under Grant 2016YFB1001300.

6. REFERENCES

- [1] J Benesty, S Makino, and J Chen, "Speech enhancement," 2005.
- [2] Philipos C Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.
- [3] Steven Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.
- [4] Jae Lim and Alan Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.
- [5] Jae S Lim and Alan V Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.
- [6] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.
- [7] Israel Cohen and Baruch Berdugo, "Speech enhancement for non-stationary noise environments," *Signal processing*, vol. 81, no. 11, pp. 2403–2418, 2001.
- [8] Israel 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.
- [9] Sunil Kamath and Philipos Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *ICASSP*. Citeseer, 2002, vol. 4, pp. 44164–44164.
- [10] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [11] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.
- [12] Xiao-Lei Zhang and Ji Wu, "Denoising deep neural networks based voice activity detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 853–857.
- [13] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2014.
- [14] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

- [15] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Global variance equalization for improving deep neural network based speech enhancement," in *Signal and Information Processing (ChinaSIP), 2014 IEEE China Summit & International Conference on*. IEEE, 2014, p-p. 71–75.
- [16] Tian Gao, Jun Du, Yong Xu, Cong Liu, Li-Rong Dai, and Chin-Hui Lee, "Improving deep neural network based speech enhancement in low snr environments," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 75–82.
- [17] Tian Gao, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Snr-based progressive learning of deep neural network for speech enhancement," *Interspeech 2016*, pp. 3713–3717, 2016.
- [18] Meng Sun, Xiongwei Zhang, Thomas Fang Zheng, et al., "Unseen noise estimation using separable deep auto encoder for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 93–104, 2016.
- [19] Andrew L Maas, Tyler M O'Neil, Awni Y Hannun, and Andrew Y Ng, "Recurrent neural network feature enhancement: The 2nd chime challenge," in *Proceedings The 2nd CHiME Workshop on Machine Listening in Multisource Environments held in conjunction with ICASSP, 2013*, pp. 79–80.
- [20] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn 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.
- [21] Szu-Wei Fu, Yu Tsao, and Xugang Lu, "Snr-aware convolutional neural network modeling for speech enhancement," *Proceedings of the Interspeech, San Francisco, CA, USA*, pp. 8–12, 2016.
- [22] Ding Liu, Paris Smaragdis, and Minje Kim, "Experiments on deep learning for speech denoising.," in *INTERSPEECH, 2014*, pp. 2685–2689.
- [23] Solomon Kullback, *Information theory and statistics*, Courier Corporation, 1997.
- [24] Fumitada Itakura, "Analysis synthesis telephony based on the maximum likelihood method," *Rep. 6th Int. Congr. Acoust.*, 1968.
- [25] Yong Xu, Jun Du, Zhen Huang, Li-Rong Dai, and Chin-Hui Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," *arXiv preprint arXiv:1703.07172*, 2017.
- [26] John S Garofolo et al., "Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST), Gaithersburgh, MD*, vol. 107, 1988.
- [27] G Hu, "100 nonspeech environmental sounds,[online] available: <http://web.cse.ohio-state.edu/pnl/corpus/hunonspeech>," *HuCorpus.html*, 2004.
- [28] Andrew Varga and Herman JM 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.
- [29] Jun Du and Qiang Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions.," in *INTER-SPEECH, 2008*, pp. 569–572.
- [30] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*. IEEE, 2001, vol. 2, pp. 749–752.
- [31] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.