A Maximum Likelihood Approach to Deep Neural Network Based Speech Dereverberation

Xin Wang*, Jun Du*, and Yannan Wang*
*University of Science and Technology of China, Hefei, Anhui, China
E-mail: wx0304@mail.ustc.edu.cn, jundu@ustc.edu.cn, wyn314@mail.ustc.edu.cn

***Abstract*—Recently, deep neural network (DNN) based speech dereverberation becomes popular with a standard minimum mean squared error (MMSE) criterion for learning the parameters. In this study, a probabilistic learning framework to estimate the DNN parameters for single-channel speech dereverberation 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 charactering the prediction error vector as a multivariate Gaussian density with a zero mean vector and an unknown covariance matrix. Our experiments demonstrate that the proposed ML-based DNN learning can achieve a better generalization capability than MMSE-based DNN learning. And all the object measures of speech quality and intelligibility are consistently improved.**

I. Introduction

Reverberation is the collection of reflected sounds from the surfaces in an enclosure like an auditorium [1]. Although appropriate reverberation can compensate the inverse square law drop-off of sound intensity in the enclosure, excessive reverberation can make the sounds run together with the loss of articulation, muddy and garbled effects. As a result, reverberation often seriously degrades speech quality and intelligibility, causing significant performance degradation in automatic speech recognition (ASR) [2] and speaker identification systems [3], [4]. Therefore, speech dereverberation becomes one of the main tasks of speech signal processing.

Many dereverberation techniques have been proposed in the past. Inverse filtering [5] is one of the commonly used single-channel speech dereverberation techniques. The dereverberated signal is estimated by convolving the reverberant signal with the inverse filter. However, in many situations, the inverse filter cannot be directly determined or accurately estimated. Furthermore, this approach assumes that the room impulse response (RIR) function is minimum-phase which is not always satisfied in real practice [6]. Wu and Wang [7] utilized a two-stage approach including inverse filtering and spectral subtraction to deal with early reverberation and late reverberation separately, which relies on an accurate estimate of the inverse filter in one microphone scenarios. The inverse filtering is only effective in a short reverberation time (RT60) range. Other studies dealt with dereverberation by exploiting the properties of speech such as modulation spectrum [8], homomorphic transformation [9] and other harmonic structures [10], [11], [12].

Recently, deep neural networks (DNNs) [13], [14] have been utilized in many speech processing areas, such as speech enhancement [15], [16], source separation [17], [18] and speech recognization [19], [20], which creates a new direction of single-channel dereverberation. Han et al. [21] also proposed the supervised learning approach based on DNN to perform speech dereverberation. They utilized DNN to learn a spectral mapping from corrupted speech to clean speech for dereverberation and denoising. Wu et al. [22] proposed a reverberation-time-aware DNN-based speech dereverberation framework to handle a wide range of reverberation times. They adopted a linear output layer, globally normalized the target features into zero mean and unit variance, and then investigated the effects of frame shift and acoustic context sizes on the dereverberated speech quality using DNN at different RT60s.

However, the objective function commonly used in DNN-based regressive tasks is mean squared error, making a strong assumption that all the feature components are equivalent. In this paper, instead of the conventional minimum mean squared error (MMSE) criterion for DNN (MMSE-DNN), we explore a maximum likelihood (ML) solution 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. We need to emphasize that the MMSE-DNN approach could be considered as a special case of the proposed ML-DNN approach with the assumption that the covariance matric is an identity matric. The evaluation on the Wall Street Journal (WSJ0) corpus shows that the proposed ML-DNN approach achieves a significantly better performance than the conventional MMSE-DNN approach in the dereverberation task. Moreover, the ML-DNN approach can also achieve a better generalization capability.

II. The Proposed ML-DNN Approach

In this study, we redefine the objective function in the probabilistic framework and adopt the maximum likelihood estimation for the parameter learning in order to further improve the generalization capability of the conventional MMSE optimization for the regression DNN, as shown in Fig. 1.

The input of DNN is the $(2\tau+1)$D-dimensional log-power spectral (LPS) feature vector of reverberant speech with an acoustic context of $2\tau+1$ neighbouring frames to consider acoustic context information. The output of DNN refers to

APSIPA ASC 2017

the D-dimensional log-power spectral (LPS) feature vector corresponding to the anechoic speech. The sigmoidal hidden units and linear output units are adopted. We assume that the DNN output vector with the input vector $\boldsymbol{x}^{(n)}$ and the DNN parameter set $\boldsymbol{W}$ at sample index n is $\hat{\boldsymbol{y}}^{(n)}(\boldsymbol{x}, \boldsymbol{W})$ while the corresponding reference vector is $\boldsymbol{y}^{(n)}$. The objective function in conventional MMSE criterion is defined as:

$$J(\boldsymbol{W}) = \frac{1}{N}\sum_{n=1}^{N}\left\|\hat{\boldsymbol{y}}^{(n)}(\boldsymbol{x}^{(n)}, \boldsymbol{W}) - \boldsymbol{y}^{(n)}\right\|_2^2 \quad (1)$$

where N represents the training set size. In the proposed ML-based DNN, the prediction error vector $\boldsymbol{e}^{(n)}$ could be defined as:

$$\boldsymbol{e}^{(n)} = \boldsymbol{y}^{(n)} - \hat{\boldsymbol{y}}^{(n)}(\boldsymbol{x}^{(n)}, \boldsymbol{W}) \quad (2)$$

which is assumed to follow a multivariate Gaussian density with a D-dimensional zero mean vector and a $D \times D$ covariance matrix $\boldsymbol{V}$:

$$p(\boldsymbol{e}^{(n)}) = \mathcal{N}(\boldsymbol{e}^{(n)}\,|\boldsymbol{0}, \boldsymbol{V}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{V}|^{\frac{1}{2}}}\exp(-\frac{1}{2}\boldsymbol{e}^{(n)^T}\boldsymbol{V}^{-1}\boldsymbol{e}^{(n)}). \quad (3)$$

If the reference vector is also a random vector, then (3) is equivalent to:

$$p(\boldsymbol{y}^{(n)}\,\big|\boldsymbol{x}^{(n)}, \boldsymbol{W}, \boldsymbol{V}) = \mathcal{N}(\boldsymbol{y}^{(n)}\,\big|\,\hat{\boldsymbol{y}}^{(n)}(\boldsymbol{x}^{(n)}, \boldsymbol{W}), \boldsymbol{V}) \quad (4)$$

which implies that the conditional distribution of $\boldsymbol{y}^{(n)}$ given $\boldsymbol{x}^{(n)}$ with the parameter set $(\boldsymbol{W}, \boldsymbol{V})$ is unimodal. Given a training set with N data pairs $(\boldsymbol{X}, \boldsymbol{Y}) = \{(\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)})\,|n = 1, 2, ...N\}$ and making the assumption that these data pairs are drawn independently from the distribution in (4), we can define the likelihood function as:

$$p(\boldsymbol{Y}\,|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{V}) = \prod_{n=1}^{N}\mathcal{N}(\boldsymbol{y}^{(n)}\,\big|\,\hat{\boldsymbol{y}}^{(n)}(\boldsymbol{x}^{(n)}, \boldsymbol{W}), \boldsymbol{V}) \quad (5)$$

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

$$\ln(p(\boldsymbol{Y}\,|\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{V})) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{V}| -$$

$$\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{y}^{(n)} - \hat{\boldsymbol{y}}^{(n)}(\boldsymbol{x}^{(n)}, \boldsymbol{W}))^T\boldsymbol{V}^{-1}(\boldsymbol{y}^{(n)} - \hat{\boldsymbol{y}}^{(n)}(\boldsymbol{x}^{(n)}, \boldsymbol{W})). \quad (6)$$

We adopt maximum likelihood criterion to alternatively optimize $\boldsymbol{W}$ and $\boldsymbol{V}$. To maximize (6) with respect to $\boldsymbol{W}$, it is equivalent to minimizing the following error function:

$$E(\boldsymbol{W}) =$$

$$\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{y}^{(n)} - \hat{\boldsymbol{y}}^{(n)}(\boldsymbol{x}^{(n)}, \boldsymbol{W}))^T\boldsymbol{V}^{-1}(\boldsymbol{y}^{(n)} - \hat{\boldsymbol{y}}^{(n)}(\boldsymbol{x}^{(n)}, \boldsymbol{W})). \quad (7)$$

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

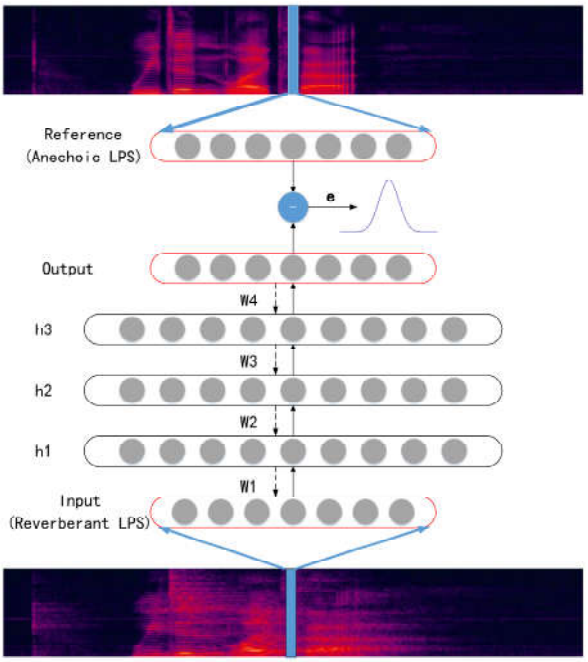


Fig. 1. The ML-DNN architecture for speech dereverberation.

Alternatively, we can also maximize (6) with respect to $\boldsymbol{V}$. The update formula can be derived as:

$$\boldsymbol{V} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{y}^{(n)} - \hat{\boldsymbol{y}}^{(n)}(\boldsymbol{x}^{(n)}, \boldsymbol{W}))(\boldsymbol{y}^{(n)} - \hat{\boldsymbol{y}}^{(n)}(\boldsymbol{x}^{(n)}, \boldsymbol{W}))^T. \quad (8)$$

To avoid the problem that the total number of parameters in the covariance matric may be very large, we use the diagonal covariance matrix in this study. The whole training procedure is summarized as Algorithm 1.

Algorithm 1 Procedure of ML-DNN training

Step 1: Initialization

Initialize the DNN parameter set $\boldsymbol{W}$. The covariance matrix $\boldsymbol{V}$ is set to an identity matrix.

Step 2: Fix $\boldsymbol{V}$ and update $\boldsymbol{W}$

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

Step 3: Fix $\boldsymbol{W}$ and update $\boldsymbol{V}$

Update $\boldsymbol{V}$ via (8).

Step 4: Go to Step 2 for L epochs

We should notice that the conventional MMSE-DNN is a special case of ML-DNN where the covariance matrix $\mathbf{V}$ in (7) is always an identity matrix, namely making a strong assumption that all the LPS components are with equivalent variances. This is the reason why MMSE optimization leads to a poor generalization capability.

APSIPA ASC 2017

III. Statistical Analysis on Prediction Errors

To verify the reasonability of our assumption that the prediction error vector $\boldsymbol{e}$ follows a multivariate Gaussian density with zero mean, we present the distributions of selected dimensions (2, 128, 257) of the prediction error vector on the cross validation set for both well-trained randomly-initialized MMSE-DNN and ML-DNN as shown 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. However, the variances are quite different, which implies that the assumption of equivalent variances in MMSE-DNN is not reasonable.

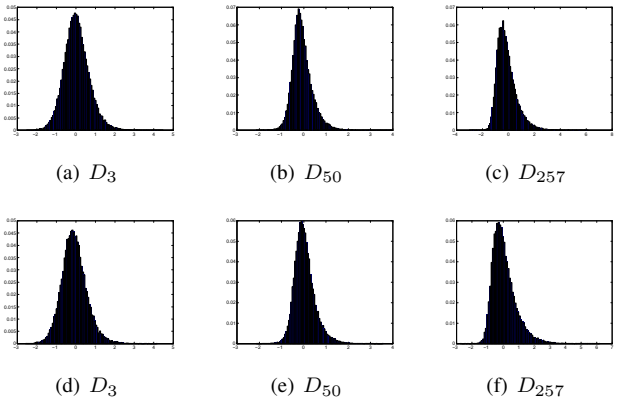
What's more, we also compare the generalization capability between MMSE-DNN initialized randomly (denoted as MMSE-DNN-InitR), ML-DNN initialized randomly (denoted as ML-DNN-InitR) and ML-DNN initialized with well-trained MMSE-DNN (denoted as ML-DNN-InitM) via the learning curves of the reconstruction loss on the cross validation (CV) set, as shown in Fig. 3. We should note that the reconstruction loss refers to the conventional mean squared error, but not the objective function proposed in (7). It is observed that the MMSE-DNN-InitR to minimize the reconstruction loss on the training data consistently generates larger errors on the cross validation set than ML-DNN-InitR which is maximizing the likelihood rather than directly minimizing the reconstruction loss in the training stage. So it is clear that ML-DNN can achieve a better generalization capability than MMSE-DNN. In addition, the learning curve of ML-DNN-InitR also shows a faster convergence than that of MMSE-DNN-InitR. Besides, ML-DNN-InitM generates smaller errors on the cross validation set than ML-DNN-InitR as shown in Fig. 3.

IV. Experiments and Results

In our experiments, the anechoic speech data were derived from the WSJ0 corpus. Twelve kinds of RIRs which were measured from three rooms with different volumes (small, medium, and large) were adopted. Then, the anechoic speech data and the RIRs were adopted for generating 7138 reverberant utterances by convolution. As for testing data, 330 anechoic utterances different from the training data were convolved with RIRs measured from three simulate rooms, whose RT60s are 0.2s, 0.5s, 0.7s respectively, to construct the whole 990 reverberant utterances.

For signal analysis, speech was sampled at 16 kHz. A 512-point DFT of each overlapping windowed frame was computed. Then 257-dimension(D=257) log-power spectral feature vectors [23] were used to train DNNs. The phase required to reconstruct waveform was directly extracted from the reverberant speech [24] and the dereverberated waveform was reconstructed from the estimated spectral magnitude and the reverberant speech phase with an overlap-add method.

All DNN configurations were fixed at three hidden layers, 2048 nodes at each hidden layer, and 7 frames of input feature expansion to consider acoustic context information. Sigmoid was used as the activation function at each hidden layer while a linear activation function was adopted at the output layer. The learning rate for the 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 40 epochs. Frequency-weighted segmental signal-to-noise ratio (fwSegSNR) [25], perceptual evaluation of speech quality (PESQ) [26], and short-time objective intelligibility (STOI) [27] were adopted to evaluate the speech signal-to-noise ratio, speech quality and speech intelligibility, respectively.



(a) D_3 (b) D_{50} (c) D_{257}

(d) D_3 (e) D_{50} (f) D_{257}

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

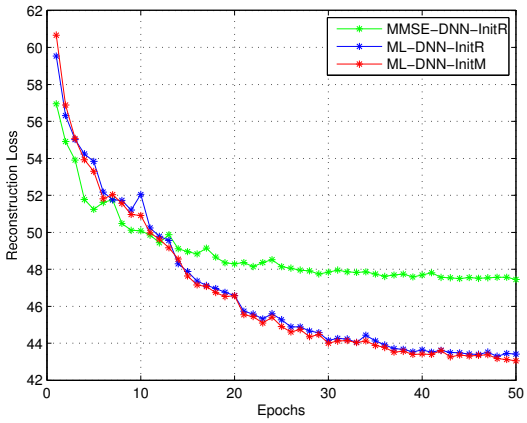


Fig. 3. The learning curve comparison of reconstruction loss among MMSE-DNN-InitR, ML-DNN-InitR and ML-DNN-InitM on the cross validation set.

Table I shows the average fwSegSNR, PESQ and STOI on the test set among MMSE-DNN-InitR, ML-DNN-InitR and ML-DNN-InitM. Clearly, the proposed ML-DNN approach yielded significant improvements over the conventional MMSE-DNN approach for all three rooms of different RT60s, e.g., a fwSegSNR gain of 0.27 dB, a PESQ gain of 0.17, a STOI gain of 0.020 in room1. In addition, the improvements of ML-DNN-InitR are smaller than those of ML-DNN-InitM. Fig. 4 shows the the spectrograms of a test utterance in room2. Clearly, ML-DNN can better restore the anechoic spectrogram than MMSE-DNN, especially in low and intermediate frequencies.

APSIPA ASC 2017

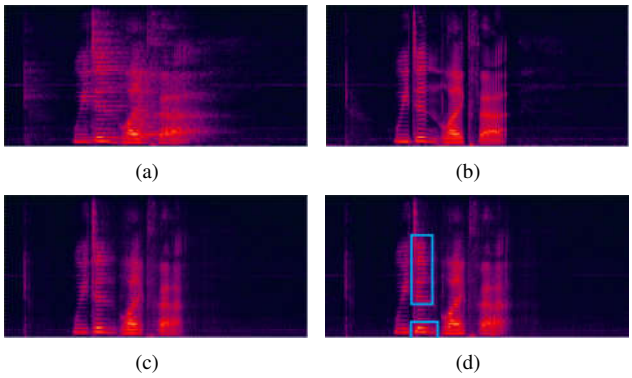


Fig. 4. Spectrograms of a test utterance in room2: (a) reverberant speech, (b) anechoic speech, (c) MMSE-DNN-InitR, (d) ML-DNN-InitR

V. Conclusions

In this study, we proposed a noval maximum likelihood approach to DNN-based speech dereverberation with a reasonable assumption that the prediction error vector of DNN follows the zero-mean Gaussian distribution. In the proposed ML-DNN, both the DNN parameters and the covariance matrix of the prediction error vector are jointly and alternatively optimized. Compared with the conventional MMSE optimization, our approach could achieve a better generalization capability and yield a better performance in the speech dereverberation task.

Acknowledgment

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.

TABLE I
Average performance comparison on the test set at rooms of different RT60s among MMSE-DNN-InitR, ML-DNN-InitR and ML-DNN-InitM.

		fwSegSNR(dB)	PESQ	STOI
MMSE-DNN-InitR	room1	12.23	2.76	0.859
	room2	11.24	2.59	0.865
	room3	9.87	2.47	0.856
ML-DNN-InitR	room1	12.50	2.93	0.879
	room2	11.49	2.73	0.885
	room3	10.04	2.59	0.871
ML-DNN-InitM	room1	12.52	2.94	0.879
	room2	11.53	2.73	0.885
	room3	10.11	2.60	0.873

References

[1] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. London, U.K.: Springer, 2010.
[2] B. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *Proc. ICASSP*, vol. 2, pp. 1259-1262, 1997.
[3] S. O. Sadjadi and J. H. L. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. ICASSP*, pp. 5448-5451, 2011.
[4] X. Zhao, Y. Wang, and D. L. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 4, pp. 836-845, 2014.
[5] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, and Lang. Process.*, vol. 36, no. 2, pp. 145-152, 1988.
[6] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, no. 1, pp. 165-169, 1979.
[7] M. Wu and D. L. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 774-784, 2006.
[8] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *Proc. ICSLP*, vol. 2, pp. 889-892, 1996.
[9] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304-1312, 1974.
[10] N. Roman and D. L. Wang, "Pitch-based monaural segregation of reverberant speech," *J. Acoust. Soc. Amer.*, vol. 120, pp. 458-469, 2006.
[11] M. Wu and D. L. Wang, "A one-microphone algorithm for reverberant speech enhancement," in *Proc. ICASSP*, pp. 844-847, 2003.
[12] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Fast estimation of a precise dereverberation filter based on speech harmonicity," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, pp. 1073-1076, 2005.
[13] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527-1554, 2006.
[14] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504-507, 2006.
[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65-68, Jan. 2014.
[16] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 1, pp. 7-19, Jan. 2015.
[17] A Narayanan and D. L. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition." *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 4, pp. 826-835, 2014.
[18] J. Du, Y.-H. Tu, Y. Xu, L.-R. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. Int. Conf. Signal Process.*, pp. 473-477, 2014.
[19] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 1, pp. 30-42, 2012.
[20] P. Golik, P. Doetsch, and H. Ney. "Cross-entropy vs. squared error training: a theoretical and experimental comparison," in *INTERSPEECH*, pp. 1756-1760, 2013.
[21] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, pp. 4628-4632, 2014.
[22] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A Reverberation-Time-Aware Approach to Speech Dereverberation Based on Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Lang. Process.*, vol. 25, no. 1, pp. 102-111, 2017.
[23] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *Proc. Interspeech*, pp. 569-572, 2008.
[24] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 4, pp. 679-681, Aug. 1982.
[25] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229-238, Jan. 2008.
[26] A.W. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Recommendation, pp. 862, 2001.
[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timeCfrequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125-2136, Sep. 2011.

APSIPA ASC 2017