An Experimental Study on Speech Enhancement Based on Deep Neural Networks

Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*—This letter presents a regression-based speech enhancement framework using deep neural networks (DNNs) with a multiple-layer deep architecture. In the DNN learning process, a large training set ensures a powerful modeling capability to estimate the complicated nonlinear mapping from observed noisy speech to desired clean signals. Acoustic context was found to improve the continuity of speech to be separated from the background noises successfully without the annoying musical artifact commonly observed in conventional speech enhancement algorithms. A series of pilot experiments were conducted under multi-condition training with more than 100 hours of simulated speech data, resulting in a good generalization capability even in mismatched testing conditions. When compared with the logarithmic minimum mean square error approach, the proposed DNN-based algorithm tends to achieve significant improvements in terms of various objective quality measures. Furthermore, in a subjective preference evaluation with 10 listeners, 76.35% of the subjects were found to prefer DNN-based enhanced speech to that obtained with other conventional technique.**

***Index Terms*—Deep neural networks, noise reduction, regression model, speech enhancement.**

I. Introduction

THE problem of enhancing noisy speech recorded by a single microphone has attracted much research effort for several decades in speech communication [1]. Many different approaches have been proposed in the literature [1]–[3] under various assumptions. Most of these techniques often can not make a good estimate of clean speech and lead to a high level of musical noise artifacts [4].

Early work on using shallow neural networks (SNNs) as nonlinear filters has also been proposed [5]–[7]. Nevertheless, the performance of the SNN model with little training data and relatively small network size is usually not satisfactory. Furthermore, gradient-based optimization, starting from random initialization, often appears to get stuck in "apparent local minima or plateaus" [8], especially when deep-layer network structures are considered. Recent insight pointed out by Hinton *et al.* [9] using a greedy layer-wise unsupervised learning procedure had resurrected the interest of the DNN and successfully applied to automatic speech recognition (ASR) and a few related tasks, outperforming the state-of-the-art systems (e.g., [10], [11]).

Other data-driven methods attempt to make a binary classification decision on time-frequency (T-F) units, such as estimating the ideal binary mask for monaural speech separation [13], however the acoustic context information of the T-F unit is not well utilized in a classification framework. In [14], DNNs were used to estimate a smoothed ideal ratio mask (IRM) in the Mel frequency domain for robust ASR.

In this study, we propose to learn the complex mapping function from noisy to clean speech with nonlinear DNN-based regression models using multi-condition training data encompassing different key factors in noisy speech, including speakers, noise types, and signal-to-noise ratios (SNRs). To our knowledge, this is one of the leading research employing a regression DNN model for speech enhancement with a large size of training data.

The rest of the letter is organized as follows. In Section II, we present the proposed DNN-based speech enhancement system. A set of evaluation experiments to assess the system performance in various DNNs configurations are provided in Section III. Finally we summarize our findings in Section IV.

II. Deep Neural Networks for Speech Enhancement

A block of the proposed speech enhancement system is illustrated in Fig. 1. In the training stage, a regression DNN model is trained from a collection of stereo data, consisting of pairs of noisy and clean speech represented by the log-power spectra features. In the enhancement stage, the well-trained DNN model is fed with the features of noisy speech in order to generate the enhanced log-power spectra features. The additional phase information is calculated from the original noisy speech. The assumption is that the phase information is not important for the human auditory perception, so only an estimate of the magnitude or power of the speech is required [7]. Finally an overlap-add method is used to synthesize the waveform of the estimated clean speech. A detailed description of the feature extraction module and the waveform reconstruction module can be found in [12].

A. Pre-training DNNs with Noisy Data

The DNN training, starting with a randomly initialized network, typically finds poor local minima [9], especially when the number of hidden layers increases. Hence, as in [17], we firstly try to learn a deep generative model of noisy log-spectra by a stacking of multiple restricted Boltzmann machines (RBMs) [8]. The left part of Fig. 2 illustrates the RBM pre-training fed with noisy data. The first one is a Gaussian-Bernoulli RBM that has one visible layer of linear variables, connected to a hidden layer. Then a pile of Bernoulli-Bernoulli RBMs can be stacked

Manuscript received July 29, 2013; revised November 06, 2013; accepted November 09, 2013. Date of publication November 14, 2013; date of current version November 20, 2013. This work was supported in part by the National Nature Science Foundation of China under Grants 61273264 and 61305002, and by the National 973 program of China under Grant 2012CB326405. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Emanuel Habets.

Y. Xu, J. Du, and L.-R. Dai are with the National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, Anhui, China (e-mail: xuyong62@mail.ustc.edu.cn; jundu@ustc.edu.cn; lrdai@ustc.edu.cn).

C.-H. Lee is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. (e-mail: chl@ece.gatech.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2013.2291240

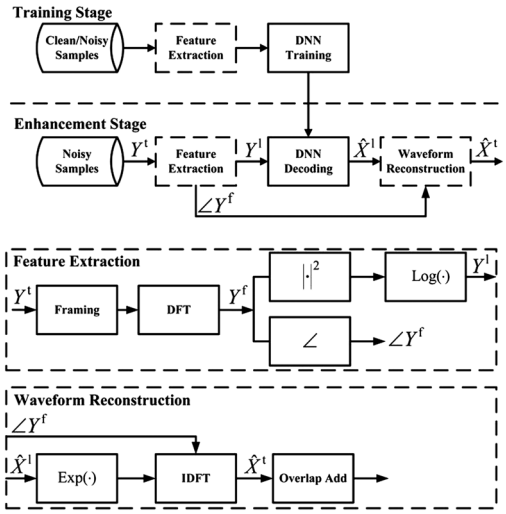

Fig. 1. A block diagram of the proposed DNN-based speech enhancement system.

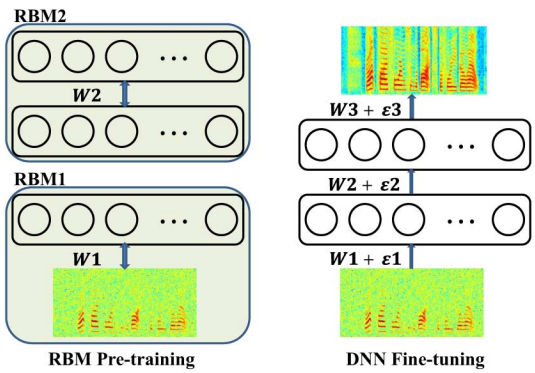

Fig. 2. Left: Illustration of the RBM pre-training that consists of only two RBMs. Right: Description of the fine-tuning procedure of DNN-based speech enhancement with the random initialized linear output layer.

behind the Gaussian-Bernoulli RBM. Afterwards, they can be trained layer-by-layer in an unsupervised greedy fashion [9]. During that, an objective criterion, called contrastive divergence (CD), is used to update the parameters of each RBM [8].

B. MMSE-based Fine-tuning

Back-propagation algorithm with the minimum mean squared error (MMSE) object function between the target and enhanced log-power spectral features is used to train the DNN. The right part of Fig. 2 describes the procedure of fine-tuning. The MMSE criterion in the log domain is more consistent with the human auditory system [6]. A stochastic gradient descent algorithm is performed in mini-batches with multiple epochs to improve learning convergence as follows,

$$E = \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}(\hat{X}_n^d(\mathbf{W}^\ell, \mathbf{b}^\ell) - X_n^d)^2. \quad (1)$$

where E is the mean squared error, $\hat{X}_n^d(\mathbf{W}^\ell, \mathbf{b}^\ell)$ and X_n^d denote the d-th enhanced and target frequency bins of the log-spectral feature at sample index n, respectively, with N representing the mini-batch size, D being the size of the log-spectral feature vector, $(\mathbf{W}^\ell, \mathbf{b}^\ell)$ denoting the weights and bias parameters to be learned at the ℓ-th layer, with L indicating the total number of hidden layers and $L+1$ representing the output layer. Then the updated estimate of the weights $\boldsymbol{W}$ and bias $\boldsymbol{b}$, with a learning rate λ, can be computed iteratively in the following:

$$(\mathbf{W}^\ell, \mathbf{b}^\ell) \leftarrow (\mathbf{W}^\ell, \mathbf{b}^\ell) - \lambda\frac{\partial E}{\partial(\mathbf{W}^\ell, \mathbf{b}^\ell)}, 1 \le \ell \le L+1. \quad (2)$$

During the derivation of the model parameters, we employ nearly no assumptions because we believe that the DNN can be used to fit the desired nonlinear mapping function. Furthermore, the independence assumption among different frequency bins, used in the conventional model-based speech enhancement methods [12], is not needed in our proposed framework. The DNN is capable of capturing the context information along the time axis (using multiple frames expansion) and along the frequency axis (using log-spectral features with full frequency bins) by concatenating them into a long input feature vector for the DNN learning.

III. Experiments and Result Analysis

All experiments below were conducted on TIMIT database [19]. As in [12], additive white Gaussian noise (AWGN) and three other types of noise recordings extracted from the Aurora2 database [18], namely *Babble*, *Restaurant* and *Street*, were used as our noise signals. All 4620 utterances from the training set of the TIMIT database [19] were added with the abovementioned four types of noise and six levels of SNR, at 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, and -5 dB, to build a multi-condition stereo training set. This resulted in a collection of about 100 hours of noisy training data (including one case of clean training data) used to train the DNN-based speech enhancement models. Another 200 randomly selected utterances from the TIMIT test set were used to construct the test set for each combination of noise types and SNR levels. Two other noise types, namely *Car* and *Exhibition*, were used for mismatch evaluation. To evaluate the performance of DNN-based speech enhancement, an improved version of the optimally modified log-spectral amplitude (OMLSA) [2], [15], [16], denoted as log-MMSE (**L-MMSE**) method, was used for performance comparison. The optimal spectral gain function of them, which minimizes the mean-square error of the log-spectra, is obtained as a weighted geometric mean of the hypothetical gains associated with the speech presence uncertainty [15].

As for signal analysis, speech waveform was down-sampled to 8KHz, and the corresponding frame length was set to 256 samples (or 32 msec) with a frame shift of 128 samples. A short-time Fourier analysis was used to compute the DFT of each overlapping windowed frame. Then 129 dimensions log-power spectra features [12] were used to train DNNs. Two objective quality measures, segmental SNR (SegSNR in dB) and log-spectral distortion (LSD in dB), were used for evaluating the quality of the enhanced speech as in [12]. In addition, perceptual evaluation of speech quality (PESQ), which has a high correlation with subjective score [20], was also used to compare system performance. In the following experiments, we only gave selective results of those three objective measures due to space limitation. Subjective listening tests would also be conducted for comparison.

The number of epoch for each layer of RBM pre-training was 20. Learning rate of pre-training was 0.0005. As for the fine-tuning, learning rate was set at 0.1 for the first 10 epochs, then decreased by 10% after every epoch. Total number of epoch was 50. The mini-batch size was set to $N = 128$. Input features of DNNs were normalized to zero mean and unit variance.

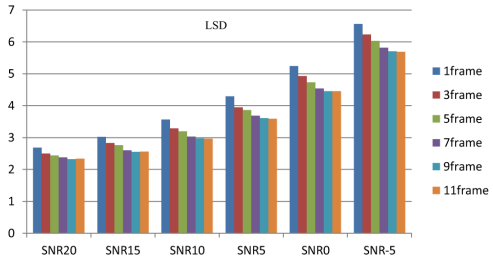

Fig. 3. Average LSD results using input with different acoustic context on the test set at different SNRs across four noise types.

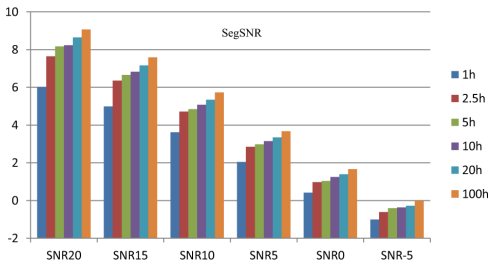

Fig. 4. Average SegSNR results using different training set size on the test set at different SNRs across four noise types.

Finally, clean condition is very special for the speech enhancement task. And for almost all speech enhancement algorithms, including L-MMSE, they do harm to the clean signal. To keep all of the information in clean utterances, a background detection operation, about whether the testing utterance was clean or not, was conducted before enhancing. It was easily implemented based on the energy and zero-crossing rate [22] of the framed utterance. With this pre-processing step, better overall results could be obtained. As the signal of clean utterances stayed unchanged after this pre-processing, the results of noise-free conditions were omitted below.

A. Evaluation of the Acoustic Context Information

Fig. 3 shows the average LSD results on the test set at different SNRs across four noise types using input features with multiple frames expansion, ranging from 1 to 11 frames at a two-frame increment. Other configurations of the DNN were $L = 3$ hidden layers, 2048 hidden units, and 100 hours training data. It is clear that the longer frames (no more than 11 frames) the DNN was fed with, the better the performance could be achieved. In addition, more acoustic context information could smooth the enhanced speech to obtain better hearing sense. However too long frames also made the DNN structure more complicated to learn in training.

B. Evaluation of the Training Set Size

Fig. 4 presents the average SegSNR results of different training set size on the test set at different SNRs across four noise types. Other configurations of the DNN were $L = 3$ hidden layers, 2048 hidden units and 11 frames expansion. Poor results were obtained if the data size was only one hour, which was almost at the same scale as that used in [7], indicating that sufficient training samples are very important to obtain a more generalized model. The performance was improved greatly when the data size was getting larger. Even up to 100 hours, the performance was not saturated.

TABLE I
AVERAGE PESQ RESULTS AMONG NOISY, L-MMSE, SNN AND DNN_L ON THE TEST SET AT DIFFERENT SNRS ACROSS FOUR NOISE TYPES. THE SUBSCRIPT L OF DNN_L REPRESENTED THE HIDDEN LAYER NUMBER

	Noisy	L-MMSE	SNN	DNN_1	DNN_2	DNN_3	DNN_4
SNR20	2.99	3.32	3.48	3.46	3.59	**3.60**	3.59
SNR15	2.65	2.99	3.26	3.24	3.35	**3.36**	**3.36**
SNR10	2.32	2.65	2.99	2.97	3.08	**3.10**	3.09
SNR5	1.98	2.30	2.68	2.65	2.76	**2.78**	**2.78**
SNR0	1.65	1.93	2.32	2.29	2.38	**2.41**	**2.41**
SNR-5	1.38	1.55	1.92	1.89	1.95	**1.97**	**1.97**
Ave	2.16	2.46	2.78	2.75	2.85	**2.87**	**2.87**

C. Overall Evaluation

For the match evaluations of noisy testing, average PESQ results among Noisy, L-MMSE, SNN, and DNN with various number of hidden layers on the test set at different SNRs across four noise types are listed in Table I. The configurations for the DNN were $L = 1$, 2, or, 3 hidden layers (denoted as DNN_L), 2048 hidden units and 11 frames of input feature expansion. As for the SNN, its configurations were $L = 1$ hidden layer, 6144 hidden layer units and 11 frames input. And the RBM pre-training was not used to initialize the weights of the SNN. The DNN and the SNN were both trained with 100 hours training data. It shows that each DNN-based method outperformed the L-MMSE method significantly indicating that the DNNs were capable of making more accurate estimation of the target speech corrupted by noise. The DNNs with more hidden layers (no more than 3 hidden layers) were demonstrated more effective and the DNN_3-based method achieved the best performance. The improvement of objective measures over the SNN which have the same number of parameters with the DNN_3 indicated that deeper architectures had a much stronger regression capability.

Table II shows the PESQ results among Noisy, L-MMSE, SNN and DNN_3 on the test set at different SNRs in mismatch environments under *Car* and *Exhibition* noises, which were both derived from Aurora2 database [18]. The *Car* noise was more stable than the *Exhibition* noise. Comparing the results of the DNN_3-based method and the L-MMSE method, the former was superior to the latter at all SNRs across two unseen noise types, especially at low SNRs and under the unstable *Exhibition* noise. These results indicated that the proposed DNN-based method had more powerful capacity to model low SNRs and unstable noise conditions. Meanwhile, the DNN-based method outperformed the SNN-based method at different SNRs across two noise types. Another two mismatch testings were also conducted, namely, (i) SNR at 7 dB which was not seen in the training set; (ii) 200 randomly selected Mandarin utterances were used as clean speech added with Babble noise at 10 dB to evaluate cross-language performance. When compared the DNN_3-based method with the L-MMSE method, PESQ of (i) was improved from 2.48 to 2.95, and PESQ of (ii) was increased from 2.20 to 2.31. Clearly the proposed DNN-based approach outperformed the L-MMSE method and the SNN-based method in all mismatched settings. Large-size multi-condition training data ensured good generalization to mismatched environments, which could be further improved using more noise types in training.

A subjective preference listening test with 10 subjects (five males and five females), under match (AGWN and Babble) and mismatch (Car) environments, comparing the DNN method with L-MMSE method, was also conducted. 36 pairs of DNN-based and L-MMSE enhanced speech utterances, from the test set for each SNR in each environment, were assigned

TABLE II
PESQ RESULTS AMONG NOISY, L-MMSE, SNN AND DNN_3 ON THE TEST SET AT DIFFERENT SNRS IN MISMATCH ENVIRONMENTS UNDER *Car* AND *Exhibition* NOISES, LABELED AS CASE A AND B, RESPECTIVELY

	Noisy		L-MMSE		SNN		DNN_3	
	A	B	A	B	A	B	A	B
SNR20	3.15	2.89	3.52	3.19	3.43	3.24	**3.58**	**3.30**
SNR15	2.81	2.55	3.23	2.85	3.19	2.96	**3.31**	**3.01**
SNR10	2.47	2.21	2.89	2.51	2.93	2.66	**3.03**	**2.69**
SNR5	2.14	1.87	2.57	2.11	2.60	2.30	**2.71**	**2.33**
SNR0	1.81	1.56	2.21	1.72	2.24	1.92	**2.35**	**1.93**
SNR-5	1.52	1.28	1.82	1.34	1.85	1.52	**1.96**	**1.54**
Ave	2.32	2.06	2.70	2.29	2.71	2.43	**2.83**	**2.47**

TABLE III
SUBJECTIVE PREFERENCE EVALUATIONS UNDER ONE MISMATCH (CAR) AND TWO MATCH ENVIRONMENTS (AGWN AND BABBLE)

	AGWN	Babble	Car	Average
DNN	86.98%	80.21%	61.86%	76.35%
L-MMSE	13.02%	19.79%	38.14%	23.65%

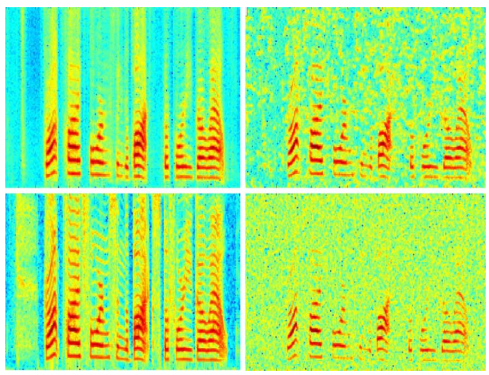

Fig. 5. Spectrograms of an utterance example with DNN enhanced (upper left), L-MMSE enhanced (upper right), original (bottom left), and noisy (bottom right) speech. Test on AGWN noise at $\mathrm{SNR} = 10$ dB.

to each listener. Table III gives the preference results given by the listeners. An average of 76.35% of the subjects preferred DNN-based to L-MMSE based enhanced speech, even under mismatched noise conditions.

Musical noise appeared in almost all traditional speech enhancement methods due to noise or SNR estimation errors leading to spurious peaks in the processed spectrum [21]. Fig. 5 displays the spectrograms of an utterance example. No musical noise was found in the DNN-enhanced spectrogram shown in the upper left panel. Furthermore, the DNN model could restore the spectrum at high frequencies buried under noise. This was not observable in the L-MMSE method (shown in the upper right panel). The DNN-enhanced spectrogram was noted to give a closer match to the original clean spectrogram (shown in the bottom left panel) than the L-MMSE enhanced version. More results and enhanced examples can be found at http://home.ustc.edu.cn/~xuyong62/demo/SE_DNN.html.

IV. CONCLUSION

In this letter, a speech enhancement framework based on the DNN, is proposed. An RBM pre-training scheme is introduced to initialize the DNN. A large training set is crucial to learn the rich structure of the DNN. Using more acoustic context information is also shown to improve the performance and make the enhanced speech less discontinuous. Multi-condition training can deal with speech enhancement of new speakers, unseen noise types, various SNR levels under different noise conditions, and even cross-language generalization. Compared with the SNN-based and L-MMSE methods, significant improvements were achieved on the TIMIT corpus. On average, 76.35% subjective preference was obtained due to the absence of musical noise in enhanced speech. This work represents our first study applying the DNN as a regression model to the speech enhancement task. In the future, we will improve the current DNN-based speech enhancement system to perform noise adaptation in real environments and to adopt objective functions relevant to auditory perception.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. ed. Boca Raton, FL, USA: CRC, 2013.

[2] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin, Germany: Springer, 2008, pp. 873–901.

[3] Y. Ephraim and D. Malah, "Speech enhancement using minimu mean square log spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, 1985.

[4] A. Hussain, M. Chetouani, S. Squartini, A. Bastari, and F. Piazza, "Nonlinear speech enhancement: An overview," in *Progress in Nonlinear Speech Processing*. Berlin, Germany: Springer, 2007, pp. 217–248.

[5] S. I. Tamura, "An analysis of a noise reduction neural network," in *Proc. ICASSP*, 1989, pp. 2001–2004.

[6] F. Xie and D. V. Compernolle, "A family of MLP based nonlinear spectral estimators for noise reduction," in *Proc. ICASSP*, 1994, pp. 53–56.

[7] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," in *Handbook of Neural Networks for Speech Processing*, S. Katagiri, Ed. Norwell, MA, USA: Artech House, 1998.

[8] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[9] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[10] G. E. Hinton, L. Deng, D. Yu, and G. E. Dahl, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[11] X. L. Zhang and J. Wu, "Denoising deep neural networks based voice activity detection," in *Proc. ICASSP*, 2013, pp. 853–857.

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

[13] Y. X. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, 2013.

[14] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2013, pp. 1520–6149.

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

[16] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, 2003.

[17] L. Deng, M. L. Seltzer, and D. Yu *et al.*, "Binary coding of speech spectrograms using a deep auto-encoder," in *Proc. Interspeech*, 2010, pp. 1692–1695.

[18] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the preformance evaluations of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, 2000, pp. 181–188.

[19] J. S. Garofolo, Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database NIST Tech Report, 1988.

[20] ITU-T, Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Int. Telecommun. Union-Telecommun. Stand. Sector 2001.

[21] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement system," in *Proc. ICASSP*, 2009, pp. 4409–4412.

[22] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1978.