This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2014.2364452, IEEE/ACM Transactions on Audio, Speech, and Language Processing

A Regression Approach to Speech Enhancement Based on Deep Neural Networks

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

***Abstract*—In contrast to the conventional minimum mean square error (MMSE) based noise reduction techniques, we propose a supervised method to enhance speech by means of finding a mapping function between noisy and clean speech signals based on deep neural networks (DNNs). In order to be able to handle a wide range of additive noises in real-world situations, a large training set that encompasses many possible combinations of speech and noise types, is first designed. A DNN architecture is then employed as a nonlinear regression function to ensure a powerful modeling capability. Several techniques have also been proposed to improve the DNN-based speech enhancement system, including global variance equalization to alleviate the over-smoothing problem of the regression model, and the dropout and noise-aware training strategies to further improve the generalization capability of DNNs to unseen noise conditions. Experimental results demonstrate that the proposed framework can achieve significant improvements in both objective and subjective measures over the conventional MMSE based technique. It is also interesting to observe that the proposed DNN approach can well suppress highly non-stationary noise, which is tough to handle in general. Furthermore, the resulting DNN model, trained with artificial synthesized data, is also effective in dealing with noisy speech data recorded in real-world scenarios without the generation of the annoying musical artifact commonly observed in conventional enhancement methods.**

***Index Terms*—Speech enhancement, noise reduction, deep neural networks, global variance equalization, non-stationary noise, noise aware training, dropout**

I. Introduction

In recent years, single-channel speech enhancement has attracted a considerable amount of research attention because of the growing challenges in many important real-world applications, including mobile speech communication, hearing aids design and robust speech recognition [1]. The goal of speech enhancement is to improve the intelligibility and quality of a noisy speech signal degraded in adverse conditions [2]. However, the performance of speech enhancement in real acoustic environments is not always satisfactory.

Numerous speech enhancement methods were developed over the past several decades. Spectral subtraction [3] subtracts an estimate of the short-term noise spectrum to produce an estimated spectrum of the clean speech. In [4], the iterative wiener filtering was presented using an all-pole model. A common problem usually encountered in these conventional methods (e.g., [3, 4]) is that the resulting enhanced speech often suffers from an annoying artifact called "musical noise" [5]. Another notable work was the minimum mean-square error (MMSE) estimator introduced by Ephraim and Malah [6]; their MMSE log-spectral amplitude estimator [7] could result in much lower residual noise without further affecting the speech quality. An optimally-modified log-spectral amplitude (OM-LSA) speech estimator and a minima controlled recursive averaging (MCRA) noise estimation approach were also presented in [8, 9]. Although these traditional MMSE-based methods are able to yield lower musical noise (e.g., [10, 11]), a trade-off in reducing speech distortion and residual noise needs to be made due to the sophisticated statistical properties of the interactions between speech and noise signals. Most of these unsupervised methods are based on either the additive nature of the background noise, or the statistical properties of the speech and noise signals. However they often fail to track non-stationary noise for real-world scenarios in unexpected acoustic conditions.

Considering the complex process of noise corruption, a nonlinear model, like the neural networks, might be suitable for modeling the mapping relationship between the noisy and clean speech signals. Early work on using shallow neural networks (SNNs) as nonlinear filters to predict the clean signal in the time or frequency domain has been proposed (e.g., [12–14]). In [15], the SNN with only one hidden layer using 160 neurons was proposed to estimate the instantaneous signal-to-noise ratios (SNRs) on the amplitude modulation spectrograms (AMS), and then the noise could be suppressed according to the estimated SNRs of different channels. However, the SNR was estimated in the limited frequency resolution with 15 channels and it was not efficient to suppress the noise type with sharp spectral peaks. Furthermore, the small network size can not fully learn the relationship between the noisy feature and the target SNRs.

In addition, random initialization of the SNNs often suffered from "apparent local minima or plateaus" [16], and the problem would get even worse for architectures incorporating more hidden layers [17]. A breakthrough for training deep architectures came in 2006 when Hinton *et al*. [18, 19]

Manuscript received March 1, 2014. Revised October 8, 2014. Accepted October 15, 2014.

This work was partially funded by the National Nature Science Foundation of China (Grant No. 61273264 and No. 61305002) and the National 973 program of China (Grant No. 2012CB326405). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. DeLiang Wang.

Yong Xu, Jun Du, and Li-Rong. Dai are with the National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China. E-mail: xuyong62@mail.ustc.edu.cn; jundu, lrdai@ustc.edu.cn. Chin-Hui Lee is with the School of Electrical and Computer Engineering, Georgia Institute of Technology. E-mail: chl@ece.gatech.edu. Jun Du is the corresponding author.

Digital Object Identifier xxxxxx

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2014.2364452, IEEE/ACM Transactions on Audio, Speech, and Language Processing

proposed a greedy layer-wise unsupervised learning algorithm. Each layer is pre-trained without supervision to learn a high level representation of its input (or the output of its previous layer). For the regression task, deep learning has been used in several speech synthesis tasks [20, 21]. In [22, 23], stacked denoising autoencoders (SDAs), as one type of the deep models, were adopted to model the relationship between clean and noisy features, and they only explored its performance on a matching test set. Deep recurrent neural networks (DRNNs) were also adopted in the feature enhancement for robust speech recognition [24, 25]. The generalization capacity of the DRNN was weak if it was trained on limited noise types [24]. [25] focused on the speech recognition evaluation for the domestic environment of CHiME corpus [26]. However, a universal speech enhancer to any noise environments is the goal in this paper.

Hence, one common problem observed for neural network based speech enhancement algorithms is the degraded performance in unseen noise conditions. A simple yet effective method to cope with the unseen noise conditions is to include many different noise types in the training set [15, 27]. Speech enhancement was formulated as a binary classification problem to estimate the ideal binary mask (IBM) in [27], and demonstrated robustness to varying background noise by training in a wide range of acoustic conditions. However, due to the binary nature of the IBM, as defined in computational auditory scene analysis (CASA) [29], it offers limited improvements to speech quality even though binary masking has been shown to improve speech intelligibility. In [27], the frequency context information of time-frequency units had not been explicitly utilized in this classification-based speech separation framework considering that the classifier was trained for each filter channel separately. However, the following work presented in [28] adopted a second DNN to capture the context information to improve the separation performance. Another smoothed ideal ratio mask (IRM) [30, 31] in the Mel frequency domain was also estimated by DNNs for robust speech recognition under seen noise types.

Recently in [32], we have proposed a regression DNN based speech enhancement framework via training a deep and wide neural network architecture using a large collection of heterogeneous training data with four noise types. It was found that the annoying musical noise artifact could be greatly reduced with the DNN-based algorithm and the enhanced speech also showed an improved speech quality both in terms of objective and subjective measures. The generalization capability of the approach was also demonstrated for new speakers, and at different SNR levels. Nonetheless the ability to handle unseen noise environments was not extensively investigated.

In this study we extend the DNN-based speech enhancement framework to handle adverse conditions and non-stationary noise types in real-world situations. In traditional speech enhancement techniques, the noise estimate is usually updated by averaging the noisy speech power spectrum using time and frequency dependent smoothing factors, which are adjusted based on the estimated speech presence probability in individual frequency bins (e.g., [8], [33]). Nonetheless, its noise tracking capacity is limited for highly non-stationary noise cases, and it tends to distort the speech component in mixed signals if it is tuned for better noise reduction. In this work, the acoustic context information, including the full frequency band and context frame expanding, is well utilized to obtain the enhanced speech with reduced discontinuity. Furthermore to improve the generalization capability we include more than 100 different noise types in designing the training set for DNN which proved to be quite effective in handling unseen noise types, especially non-stationary noise components.

Three strategies are also proposed to further improve the quality of enhanced speech and generalization capability of DNNs. First, an equalization between the global variance (GV) of the enhanced features and the reference clean speech features is proposed to alleviate the over-smoothing issue in DNN-based speech enhancement system. The second technique, called dropout, is a recently proposed strategy for training neural networks on data sets where over-fitting may be a concern [34]. While this method was not designed for noise reduction, it was demonstrated [35] to be useful for noise robust speech recognition and we successfully apply it to a DNN as a regression model to produce a network that has a good generalization ability to variabilities in the input. Finally, noise aware training (NAT), first proposed in [35], is adopted to improve performance.

The rest of the paper is organized as follows. We first give an overview of our proposed speech enhancement system in Section II. Section III elaborates the basic DNN training procedure and several strategies for further improvements. A series of experiments to assess the system performance are presented in Section IV. Finally we summarize our findings in Section V.

II. System Overview

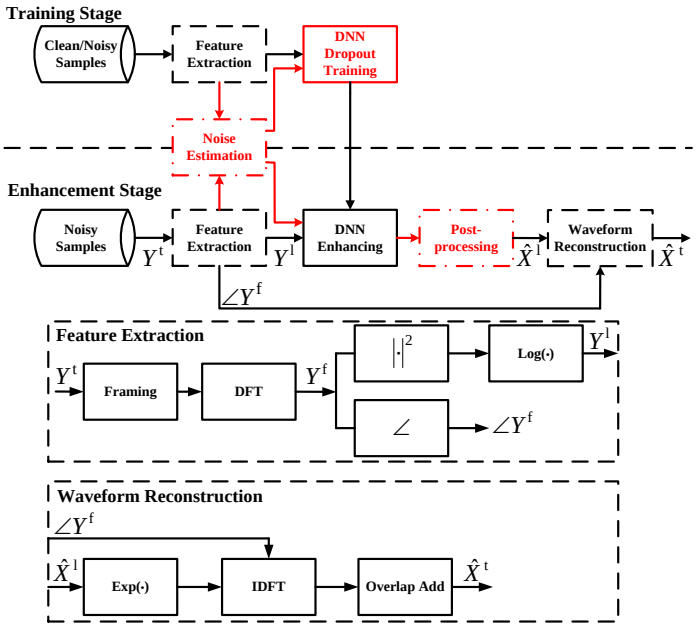


Fig. 1. A block diagram of the proposed DNN-based speech enhancement system.

A block diagram of the proposed speech enhancement

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2014.2364452, IEEE/ACM Transactions on Audio, Speech, and Language Processing

framework is illustrated in Fig. 1. A DNN is adopted as the mapping function from noisy to clean speech features. Our baseline system [32] is constructed in two stages. In the training stage, a DNN-based regression model was trained using the log-power spectral features from pairs of noisy and clean speech data. The log-power spectral features is adopted [40] since it is thought to offer perceptually relevant parameters (e.g., [13], [14]). Therefore, short-time Fourier analysis is first applied to the input signal, computing the discrete Fourier transform (DFT) of each overlapping windowed frame. Then the log-power spectra are calculated.

In the enhancement stage, the noisy speech features are processed by the well-trained DNN model to predict the clean speech features. After we obtain the estimated log-power spectral features of clean speech, $\hat{X}^{\mathrm{l}}(d)$, the reconstructed spectrum $\hat{X}^{\mathrm{f}}(d)$ is given by:

$$\hat{X}^{\mathrm{f}}(d) = \exp\{\hat{X}^{\mathrm{l}}(d)/2\}\exp\{j\angle Y^{\mathrm{f}}(d)\}. \quad (1)$$

where $\angle Y^{\mathrm{f}}(d)$ denotes d^{th} dimension phase of the noisy speech. Although phase information is important in human speech recognition [39], here, phase was extracted directly from the noisy signal considering that our ears are insensitive to small phase distortions or global spectral shifts [14]. However, we also pointed that the clean and noisy phases are quite different at low SNRs, unfortunately, it is harder to estimate the phase. Hence, only an estimate of the magnitude of clean speech is required here. A frame of speech signal, $\hat{\mathbf{X}}^{\mathrm{t}}$, can now be derived from inverse DFT (IDFT) of the current frame spectrum. Finally, an overlap-add method, as in [40], is used to synthesize the waveform of the whole utterance. For the sake of simplicity, we will omit the superscripts of Y^{t}, Y^{l}, Y^{f}, $\hat{X}^{\mathrm{t}}$, $\hat{X}^{\mathrm{l}}$ and $\hat{X}^{\mathrm{f}}$ in following sections.

Another two modules, namely noise estimation for noise-aware training and post-processing with global variance equalization, shown in the red dashed boxes of the system block diagram in Fig. 1, are proposed to improve the overall performance of the proposed DNN-based speech enhancement system. The dropout training strategy is also adopted to improve the generalization capacity of DNNs. Details of the proposed improvements are presented in Sec. III next.

III. DNN-BASED SPEECH ENHANCEMENT

In the following subsections, we first describe the basic DNN training procedure used in [32] and then propose several techniques to improve the baseline DNN system so that the quality of the enhanced speech in matched noise conditions can be maintained while the generalization capability to unseen noise can be increased.

A. Basic DNN Training

The architecture adopted here is a feed-forward neural network with many levels of non-linearities [51] allowing them to represent a highly non-linear regression function that maps noisy speech features to clean speech features. Note that the features are all normalized to zero mean and unit variance. The training of DNN as a regression model consists

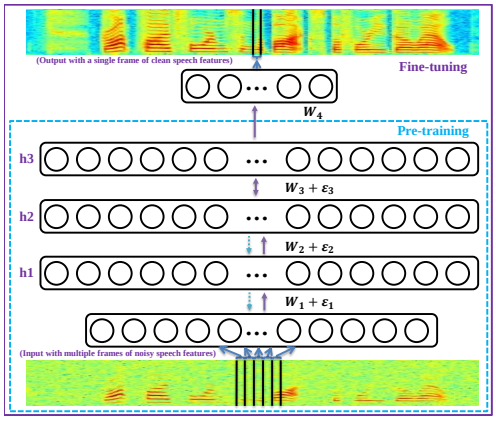


Fig. 2. Illustration of the basic DNN training procedure.

of an unsupervised pre-training part and a supervised fine-tuning part as illustrated in Fig. 2. The type of the hidden units is sigmoid, and the output unit is linear. To avoid getting stuck in local minima when training deep networks [17, 19, 42], we first pre-train a deep generative model with the normalized log-power spectra of noisy speech by stacking multiple restricted Boltzmann machines (RBMs) [16] as shown in the dashed blue box of Fig. 2. Since the input feature vectors are of real-valued in our DNNs, the first RBM in Fig. 2 is a Gaussian-Bernoulli RBM that has one visible layer of Gaussian variables, connected to a hidden binary layer. Then multiple Bernoulli-Bernoulli RBMs can be stacked on top of the Gaussian-Bernoulli RBM. They are trained layer-by-layer in an unsupervised greedy fashion to maximize the likelihood over training samples [19]. During that procedure, an objective criterion, called contrastive divergence (CD), is used to update the parameters of each RBM [16, 18].

Then the back-propagation algorithm with the MMSE-based object function between the normalized log-power spectral features of the estimated and the reference clean speech is adopted to train the DNN. In contrast to pre-training for initializing the parameters in the first several hidden layers, the fine-tuning part shown in Fig. 2 performs supervised training of all the parameters in the network. The MMSE criterion in the log-power spectral domain has shown a consistency with the human auditory system [13]. A mini-batch stochastic gradient descent algorithm is used to improve the following error function,

$$Er = \frac{1}{N}\sum_{n=1}^{N}\|\hat{\mathbf{X}}_n(\mathbf{Y}_{n-\tau}^{n+\tau}, \mathbf{W}, \mathbf{b}) - \mathbf{X}_n\|_2^2. \quad (2)$$

where Er is the mean squared error, $\hat{\mathbf{X}}_n(\mathbf{Y}_{n-\tau}^{n+\tau}, \mathbf{W}, \mathbf{b})$ and $\mathbf{X}_n$ denote the estimated and reference normalized log-spectral features at sample index n, respectively, with N representing the mini-batch size, $\mathbf{Y}_{n-\tau}^{n+\tau}$ being the noisy log-spectral feature vector where the window size of context is $2*\tau+1$, $(\mathbf{W}, \mathbf{b})$ denoting the weight and bias parameters to be learned. Then the updated estimate of $\mathbf{W}^{\ell}$ and $\mathbf{b}^{\ell}$ in the ℓ-th layer, with a

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2014.2364452, IEEE/ACM Transactions on Audio, Speech, and Language Processing

learning rate λ, can be computed iteratively in the following:

$$\Delta(\mathbf{W}_{n+1}^{\ell}, \mathbf{b}_{n+1}^{\ell}) = -\lambda \frac{\partial Er}{\partial(\mathbf{W}_{n}^{\ell}, \mathbf{b}_{n}^{\ell})} \quad (3)$$
$$-\kappa\lambda(\mathbf{W}_{n}^{\ell}, \mathbf{b}_{n}^{\ell}) + \omega\Delta(\mathbf{W}_{n}^{\ell}, \mathbf{b}_{n}^{\ell}),\ 1 \le \ell \le L+1.$$

where L denoted the total number of hidden layers and $L+1$ represented the output layer. κ is the weight decay coefficient. And ω is the momentum.

During learning, a DNN is used to learn the mapping function; no assumptions are made about the relationship of noisy speech with clean speech. It can automatically learn the complicated relationship to separate speech from the noisy signals given the sufficient training samples. Furthermore, as shown in Fig. 2, the DNN could capture the acoustic context information along the time axis (using multiple frames of noisy speech as input) and along the frequency axis (using full-band spectrum information) by concatenating them into a long input feature vector for DNN learning while the independence assumption among different dimensions was a common practice in the Gaussian mixture model to reduce computation complexity as in [40].

B. Post-processing with Global Variance Equalization

One of the residual error problems, namely over-smoothing, causes a muffling effect on the estimated clean speech when compared with reference clean speech. An equalization between the global variance of the estimated and reference clean speech features is proposed to alleviate this problem. Global variance equalization here can be considered as a simple type of histogram equalization (HEQ), which plays a key role in density matching [53]. In [43], it is demonstrated that the use of global variance information could significantly improve the subjective score in a voice conversion task.

The global variance of the estimated clean speech features is defined as:

$$GV(d) = \frac{1}{M}\sum_{n=1}^{M}(\hat{X}_n(d) - \frac{1}{M}\sum_{n=1}^{M}\hat{X}_n(d))^2. \quad (4)$$

where $\hat{X}_n(d)$ is the d-th component of a DNN output vector at the n-th frame and M is the total number of speech frames in the training set. The global variance of the normalized reference clean speech features can be calculated in a similar way. Meanwhile, a dimension-independent global variance can be computed as follows:

$$GV = \frac{1}{M*D}\sum_{n=1}^{M}\sum_{d=1}^{D}(\hat{X}_n(d) - \frac{1}{M*D}\sum_{n=1}^{M}\sum_{d=1}^{D}\hat{X}_n(d))^2. \quad (5)$$

Fig. 3 shows the global variances of the estimated and reference normalized log-power spectra of clean speech across different frequency bins. It can be observed that the global variances of the estimated clean speech features were smaller than those of the reference clean speech features, indicating that the spectra of estimated clean speech were smoothed. Moreover, this over-smoothing problem would get even worse for the lower SNR case. Fig. 4 presents the spectrograms of an utterance with additive white Gaussian noise (*AWGN*) at SNR = 0dB: DNN model trained with 104 noise types enhanced (left), clean (middle) and noisy (right) speech. A severe over-smoothing phenomenon could be observed. The formant peaks were suppressed, especially in the high frequency band which leads to muffled speech.

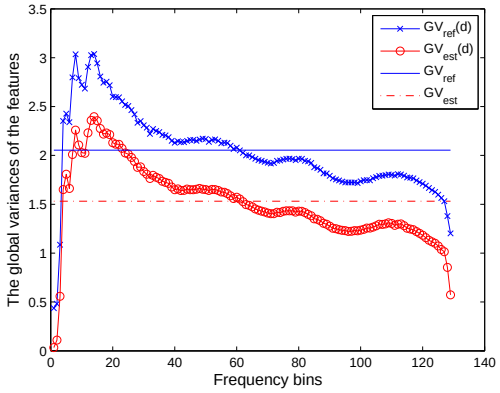


Fig. 3. The dimension-dependent and dimension-independent global variances of the reference and estimated clean speech features on the training set.

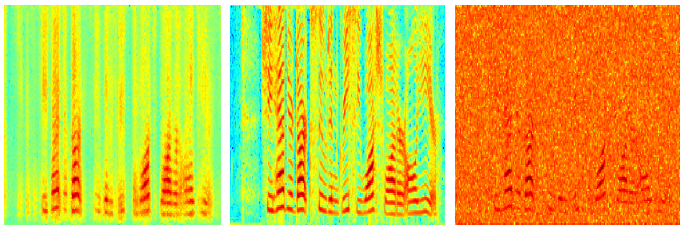


Fig. 4. Spectrograms of an utterance tested with *AWGN* at 0dB SNR: for the DNN estimated (left), the clean (middle) and the noisy (right) speech.

To address the over-smoothing problem, a global equalization factor $\alpha(d)$ is defined as follows:

$$\alpha(d) = \sqrt{\frac{GV_{ref}(d)}{GV_{est}(d)}}. \quad (6)$$

Where $GV_{ref}(d)$ and $GV_{est}(d)$ represented the d-th dimension of the global variance of the reference features and the estimation features, respectively. Furthermore, a dimension-independent global equalization factor β can be defined as:

$$\beta = \sqrt{\frac{GV_{ref}}{GV_{est}}}. \quad (7)$$

Where GV_{ref} and GV_{est} represented the dimension-independent global variance of the reference features and the estimation features, respectively.

As the input features of the DNN were normalized to zero mean and unit variance. The output of DNN $\hat{X}(d)$ should be transformed back as follows:

$$\hat{X}'(d) = \hat{X}(d) * v(d) + m(d), \quad (8)$$

where $m(d)$ and $v(d)$ are the d-th component of the mean and variance of the input noisy speech features, respectively. Then the equalization factor η could be used to lift the variance of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2014.2364452, IEEE/ACM Transactions on Audio, Speech, and Language Processing

this reconstruction signal as the post-processing:

$$\hat{X}''(d) = \hat{X}(d) * \eta * v(d) + m(d), \tag{9}$$

where η could be replaced by α or β defined in Eqs. (6)-(7). Since the DNN output $\hat{X}(d)$ was in the normalized log-power spectrum domain, the multiplicative factor η (with its options α and β) was just operated as a exponential factor in the linear spectrum domain. And this exponential factor could effectively sharpen the formant peaks of the recovered speech and suppress the residual noise simultaneously, which could significantly improve the overall listening quality demonstrated in Sec. IV.

C. Dropout Training

One of the challenges in designing DNN-based speech enhancement systems is to address possible mismatches between the training and testing conditions, caused by different SNR levels, speaker variabilities, noise types, etc. As for the first two factors, we have partially tackled them in [32]. However the mismatch in noise types is the most difficult one as there are many kinds of complicated noise environments in the real world. In this work, frame-wise DNN training fed with noisy speech features with many different noise types might be a possible solution.

To better address those mismatch issues, a strategy called "dropout" [34] could also be adopted to further improve the generalization capability of the DNN. In the DNN training, dropout randomly omits a certain percentage (e.g., ρ) of the neurons in the input and each hidden layer during each presentation of the sample for each training sample, which can be treated as model averaging to avoid the over-fitting problem. This prevents complex co-adaptations wherein the activations of multiple nodes are highly correlated [34]. Since the frequency bins of each sample are randomly omitted, and each higher-layer neuron also gets input from a random collection of the lower-layer neurons, it indeed destroys the specific relationship in noisy speech by introducing perturbations.

This operation might cause the performance degradation for matching noise types, while it could improve the robustness in mismatched cases, especially for non-stationary noises not seen in the training data. At the enhancement stage, the DNN discounts all the weights involved in the dropout training by $(1-\rho)$, instead of using a random combination of the neurons at each hidden layer [35].

D. Noise-aware Training (NAT)

In conventional speech enhancement, the noise and clean speech spectra are dynamically estimated using previous information under some model assumptions. For instance in OM-LSA approach (e.g., [8, 9]), its noise estimate is obtained by averaging previous several frames of power spectra of noisy speech, using a time-varying frequency-dependent smoothing parameter that is adjusted by the signal presence probability [9]. However, the relationship between the clean speech and noise signals is non-linear and complicated. It is therefore difficult to estimate the clean speech spectra with simple model assumptions, especially for non-stationary noises.

On the other hand, the noise information of each utterance was not specifically utilized in the basic DNN training. To enable this noise awareness, the DNN is fed with the noisy speech samples augmented with an estimate of the noise. In this way, the DNN can use additional on-line noise information to better predict the clean speech. Also the estimated noise could be regarded as a specific code for adaptation, like a speaker code in speaker adaptation [37]. Here the input vector of the DNN is similar to what was adopted in [35] with a noise estimate appended:

$$\mathbf{V}_n = [\mathbf{Y}_{n-\tau}, ..., \mathbf{Y}_{n-1}, \mathbf{Y}_n, \mathbf{Y}_{n+1}, ..., \mathbf{Y}_{n+\tau}, \hat{\mathbf{Z}}_n] \tag{10}$$

$$\hat{\mathbf{Z}}_n = \frac{1}{T}\sum_{t=1}^{T} \mathbf{Y}_t \tag{11}$$

where $\mathbf{Y}_n$ represents the log-power spectral feature vector of the current noisy speech frame n, the window size of context here is $2*\tau+1$, and the noise $\hat{\mathbf{Z}}_n$ is fixed over the utterance and estimated using the first T frames. Although this noise estimator is simple and not always efficient in robust speech recognition task [36], its effect in the speech enhancement task is not evaluated. Furthermore, the dropout to the estimated noise $\hat{\mathbf{Z}}_n$ spliced in the input layer of DNNs could compensate for the possible variability of the noise spectrum in other frames of the current utterance.

IV. Experimental Results and Analysis

In [32], only four noise types, namely *AWGN*, *Babble*, *Restaurant* and *Street*, from the Aurora2 database [44] were used as the noise signals for synthesizing the noisy speech training samples. In this study we increased the number of noise types to 104 with another 100 environmental noises [49][1]. The spectrograms of these 104 noise types were presented in Fig. 5. The clean speech data was still derived from the TIMIT database [45]. All 4620 utterances from the training set of the TIMIT database were corrupted with the abovementioned 104 noise types at six levels of SNR, i.e., 20dB, 15dB, 10dB, 5dB, 0dB, and -5dB, to build a multi-condition training set, consisting of pairs of clean and noisy speech utterances. This resulted in a collection of about 2500 hours of noisy training data (including one condition of clean training data) used to train the DNN.

We randomly select part of them to construct a 100-hour subset and a 625-hour training subset. 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. As we only conduct the evaluation of mismatched noise types in this paper, 15 other unseen noise types[2], from

[1]The 104 noise types for training are N1-N17: Crowd noise; N18-N29: Machine noise; N30-N43: Alarm and siren; N44-N46: Traffic and car noise; N47-N55: Animal sound; N56-N69: Water sound; N70-N78: Wind; N79-N82: Bell; N83-N85: Cough; N86: Clap; N87: Snore; N88: Click; N88-N90: Laugh; N91-N92: Yawn; N93: Cry; N94: Shower; N95: Tooth brushing; N96-N97: Footsteps; N98: Door moving; N99-N100: Phone dialing. To compare with the results of [32], N101: AWGN, N102: Babble, N103: Restaurant, N104: Street, were also used.

[2]The 15 unseen environment noises for evaluation are Exhibition, Car, Buccaneer1, Buccaneer2, Destroyer engine, Destroyer ops, F16, Factory1, HF channel, Leopard, Machine gun, and Pink. The first two noises are from the Aurora2 database and the others are collected from the NOISEX-92 corpus.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2014.2364452, IEEE/ACM Transactions on Audio, Speech, and Language Processing

the Aurora2 database [44] and the NOISEX-92 corpus [38], were used for testing. It should be noted that most of the following experiments were only evaluated on three typical noise types, namely, *Exhibition*, *Destroyer engine* and *HF channel*, and the overall evaluation on the whole 15 unseen noise types was given in Sec. IV-D. An improved version of OM-LSA [8, 9], denoted as **LogMMSE**, was used for performance comparison with our DNN approach.

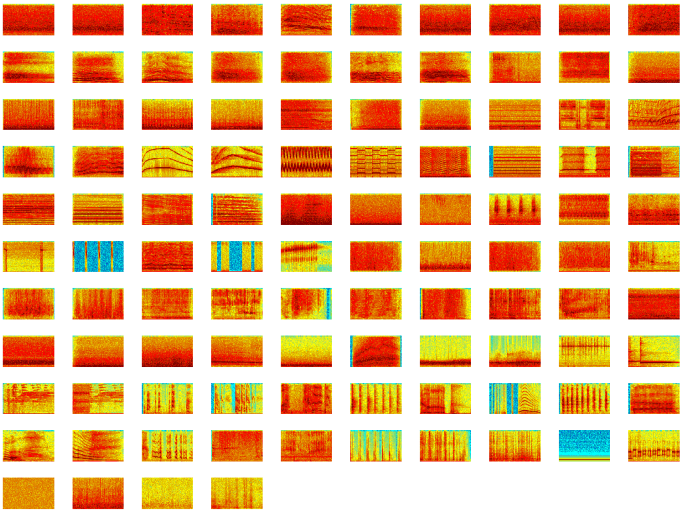


Fig. 5. Spectrograms of 104 noise types which were used as the noise signals for synthesizing the noisy speech training samples.

All the clean speech and noise waveforms were down-sampled to 8KHz. The frame length is 32 msec (256 samples) while the frame shift is 16 msec (128 samples). Then the dimension of the log-power spectral feature vector is 129. Perceptual evaluation of speech quality (PESQ) [46], which has a high correlation with subjective evaluation scores [46], was mostly used as a compressive objective measure. PESQ is calculated by comparing the enhanced speech with the clean reference speech, and it ranges from -0.5 to 4.5. We will only report limited evaluation results based on other objective measures, such as Short-Time Objective Intelligibility (STOI) score [47], segmental SNR (SSNR, in dB) [40] and log-spectral distortion (LSD, in dB) [40]. All of them are obtained by comparing the enhanced speech with the clean reference speech. STOI is highly relevant to the human speech intelligibility score ranging from 0 to 1. SSNR denotes the degree noise reduction, while LSD represents the speech distortion. Subjective measures such as analysis of spectrograms and informal listening tests will also be conducted for comparison.

The number of epoch for the RBM pre-training in each layer was 20. The learning rate of pre-training was set as 0.0005. As for the fine-tuning of the baseline, the learning rate was set to 0.1 for the first 10 epochs, then decreased by 10% after each subsequent epoch. The momentum rate ω is set to 0.9. The total number of epoch at this stage was 50. The mini-batch size N was set to 128. The weight decay coefficient κ in Eq. (3) was 0.00001. As for the back-propagation algorithm improved by the dropout regularization, the corruption levels are 0.1 for the input layer and and 0.2 at each hidden layer, respectively. The learning rate of dropout was 1. The initial momentum rate of dropout is 0.5 and then the rate increases to 0.9 in the first 10 epochs, after which it is kept as 0.9. The first $T = 6$ frames of each utterance were used for a noise estimate in NAT. Mean and variance normalization was applied to the input and target feature vectors of the DNN, so the dynamic range of the log-power spectra could be compressed to make them amenable to the back-propagation training.

As in [32], the clean speech condition was specially treated. Almost all speech enhancement methods have the side effect on the detail of the clean speech spectra. Fortunately, this has little impact on human listening. Nonetheless, to ensure that the clean signal is not distorted, a detection operation for clean speech condition, was conducted. It was easily implemented according to the energy and zero-crossing rate [48] information. With this simple step, better overall results could be obtained. So the results of the clean condition are omitted in the remainder of the paper.

In the followings, we first tuned the parameters of different DNN configurations, compared the proposed normalized clean Log-power spectra with the mask-based training targets and verified the different initialization schemes. Then the evaluations of the proposed strategies demonstrated their effectiveness to improve the generalization capacity to unseen noises. The suppression against highly non-stationary noise was also found. Finally, overall performance comparisons on 15 unseen noises and on real-world noises between the proposed method and the LogMMSE method were given.

A. Evaluations of Different DNN Configurations for Unseen Noise Environments

1) The number of noise types: In [32], we had trained a DNN model using 100 hours of noisy speech data with only four noise types, namely *AWGN*, *Babble*, *Restaurant* and *Street* noises. To improve the generalization capability of the DNN in mismatched noise conditions, we used additional 100 noise types provided in [49] to train a DNN with the same amount of training data and network configurations as in [32], namely, 11-frame expansion, 3 hidden layers, and 2048 hidden units for each hidden layer. Table I lists a performance comparison of different number of noise types using PESQ and LSD measures on the test set at different SNR levels of three unseen noise environments, namely *Exhibition*, *Destroyer engine* and *HF channel*. It was clear that the model trained with 104 noise types could achieve a better performance under the same amount of training data and DNN configurations. For example for the difficult *HF channel* case shown in Table I , the average LSD over six different SNR levels (from -5dB to 20dB) of three unseen noise types was reduced from 6.90 to 5.73. And the average PESQ was improved from 2.43 to 2.60.

2) The depth of DNN: In Table II, we compare average PESQ results at different SNRs across the abovementioned three unseen noise types using the conventional shallow neural networks (SNNs) with only one hidden layer and DNN_L. Here L denoted the number of hidden layers. The chosen DNN configurations were 11-frame expansion, 2048 hidden units in each hidden layer, and 100 hours of training data with 104 noise types. Two types of SNNs, namely SNN1 with 512 hidden units and SNN2 with 6144 (=2048*3) hidden units, both

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2014.2364452, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE I
PESQ AND LSD COMPARISON BETWEEN MODELS TRAINED WITH FOUR NOISE TYPES AND 104 NOISE TYPES ON THE TEST SET AT DIFFERENT SNRS OF THREE UNSEEN NOISE ENVIRONMENTS.

	PESQ		LSD	
	4 noise types	104 noise types	4 noise types	104 noise types
SNR20	3.23	**3.39**	2.72	**2.30**
SNR15	2.93	**3.10**	3.53	**2.90**
SNR10	2.53	**2.80**	4.99	**4.08**
SNR5	2.27	**2.46**	7.12	**5.87**
SNR0	1.92	**2.10**	9.89	**8.23**
SNR-5	1.59	**1.74**	13.14	**11.00**
Ave	2.43	**2.60**	6.90	**5.73**

with 11-frame input, were compared. SNN2 (PESQ=2.57) was shown to be superior to SNN1 (PESQ=2.48), indicating that the speech component could be separated more easily from its mixed signal with wider hidden layer in the SNN. It was also observed that DNNs with more than one hidden layer were demonstrated to be more effective and DNN_3 achieved the best performance at PESQ=2.6. The improvement of DNN_3 over SNN2 which had the same number of parameters with the DNN_3 indicated that deeper neural network architectures had a better regression capability.

TABLE II
AVERAGE PESQ RESULTS AMONG SNNS AND DNN_L ON THE TEST SET AT DIFFERENT SNRS ACROSS THE SAME THREE UNSEEN NOISE TYPES.

	Noisy	SNN1	SNN2	DNN_1	DNN_2	DNN_3	DNN_4
SNR20	2.88	3.23	3.35	3.34	3.38	**3.39**	3.37
SNR15	2.55	2.97	3.06	3.06	**3.10**	**3.10**	3.09
SNR10	2.22	2.67	2.76	2.76	2.79	**2.80**	2.78
SNR5	1.90	2.35	2.44	2.43	**2.46**	**2.46**	2.44
SNR0	1.61	2.00	2.07	2.07	**2.10**	**2.10**	2.09
SNR-5	1.37	1.66	1.72	1.72	1.74	1.74	**1.75**
Ave	2.09	2.48	2.57	2.56	2.59	**2.60**	2.59

3) The length of acoustic context: In Fig. 6 we show the average PESQ results on the test set at different SNRs across the abovementioned three mismatched noise types using input features with different size of context expansion, ranging from 1 to 13 frames at a selective frame number increment. Other configurations of the DNN were 3 hidden layers, 2048 units at each hidden layer, and 100 hours of training data with 104 noise types. We could see the longer context used (no more than 11 frames), the better the performance. In addition, more acoustic context information could reduce the discontinuity of the estimated clean speech signals to obtain a better listening quality. However using too many frames in context also degraded the performance as irrelevant information with the current frame was included.

4) The size of training set: Fig. 7 compares the average PESQ results of different training set size with 104 noise types on the test set across the three mismatched noise types at different SNRs. DNNs were configured with 3 hidden layers, 2048 units in each hidden layer and 11-frame context expansion. Poor results were obtained if the data size was only one hour, indicating that sufficient training samples are critical to obtain models with a good generalization capability. There was a big jump of performance when the training set size increased to 5 hours. The performance was improved monotonically when the data size increased until to 100 hours. The DNN trained with 625 hours data was slightly better than the DNN

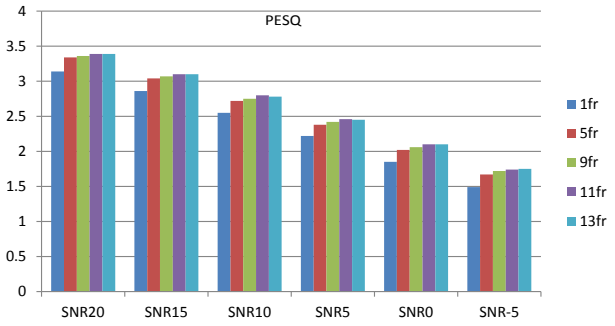


Fig. 6. Average PESQ results using different acoustic context on the test set across three unseen noise types at different SNRs.

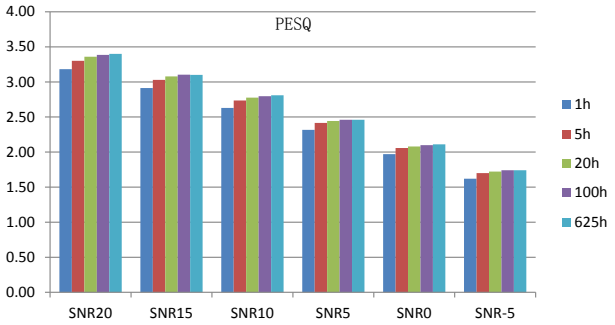


Fig. 7. Average PESQ results using different training set size with 104 noise types on the test set across three unseen noise types at different SNRs.

trained with 100 hours data. The reason is that only 4-hour clean TIMIT corpus [45] data and limited noise samples were used to construct the multi-condition training data. The data redundancy will be more severe when increasing the training data to 625 hours. We pointed out that the richness of the clean speech samples and the noise samples are the two crucial aspects to improve the generalization capacity of DNNs.

5) Comparing with the mask based training targets: The ideal ratio mask (IRM) and the short-time Fourier transform spectral mask (FFT-MASK) were well defined in [50]. And they were demonstrated to be superior to other existing training targets [50], such as, ideal binary mask (IBM), target binary mask (TBM), the short-time Fourier transform spectral magnitude (FFT-MAG), etc. Following the practice in [50], the output type of the DNN for predicting IRM is sigmoid, while the output type for FFT-MASK is linear. Hence, Table III presented the PESQ results among the proposed normalized clean log-power spectra, denoted as (a), IRM, denoted as (b) and FFT-MASK, denoted as (c) on the test set at different SNRs of three unseen noise environments. The proposed normalized clean Log-power spectra target was better than IRM and FFT-MASK at all conditions in our experimental setup. IRM and FFT-MASK got the almost the same performance. It should be noted that the proposed clean Log-power spectra normalized to mean zero and unit variance is crucial, which is different from the FFT-MAG with the Log compression followed by the percent normalization. And the MVN is better than the percent normalization used in [50], because the calculated mean and variance is more robust than the minimum and maximum value used in the percent

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2014.2364452, IEEE/ACM Transactions on Audio, Speech, and Language Processing

normalization. As for the IRM, it assumes the independence between the noise and the speech in its testing phase, although it can restrict the dynamical value range to [0, 1] in the training phase. Another main difference is that a set of features, such as, amplitude modulation spectrogram (AMS), mel-frequency cepstral coefficients (MFCC), etc, were adopted as the input of DNNs in [50]. However, the normalized noisy Log-power spectra was directly used as the input in this paper to predict the clean Log-power spectra.

TABLE III
PESQ RESULTS OF USING DIFFERENT TRAINING TARGETS: THE PROPOSED NORMALIZED CLEAN LOG-POWER SPECTRA, DENOTED AS (A), IRM, DENOTED AS (B) AND FFT-MASK, DENOTED AS (C) ON THE TEST SET AT DIFFERENT SNRS OF THREE UNSEEN NOISE ENVIRONMENTS.

	Exhibition			Destroyer engine			HF channel		
	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
SNR20	**3.37**	3.26	3.24	**3.51**	3.43	3.41	**3.27**	3.11	3.11
SNR15	**3.09**	2.96	2.96	**3.27**	3.14	3.16	**2.95**	2.77	2.79
SNR10	**2.78**	2.63	2.66	**3.00**	2.83	2.88	**2.61**	2.43	2.47
SNR5	**2.41**	2.27	2.30	**2.69**	2.50	2.57	**2.28**	2.10	2.14
SNR0	**1.99**	1.88	1.90	**2.35**	2.15	2.22	**1.96**	1.80	1.83
SNR-5	**1.57**	1.51	1.51	**2.00**	1.81	1.87	**1.65**	1.56	1.55
Ave	**2.53**	2.42	2.43	**2.80**	2.64	2.69	**2.45**	2.30	2.31

6) Comparing RBM pre-training with the random initialization: Table IV presented the PESQ, LSD and SSNR results of using RBM pre-training with 100 hours training data and random initialization on the test set at different SNRs of three unseen noise environments. RBM pre-training was slightly better than the random initialization at low SNR conditions where the mapping function is more complicated. Noted that the training data was 100 hours here. With more training data, the back-propagation algorithm could not be stuck in the local optimum. And the RBM pre-training will be more beneficial when the training data is insufficient.

TABLE IV
PESQ, LSD AND SSNR RESULTS OF USING RBM PRE-TRAINING WITH 100 HOURS TRAINING DATA AND RANDOM INITIALIZATION ON THE TEST SET AT DIFFERENT SNRS OF THREE UNSEEN NOISE ENVIRONMENTS.

	RBM			Random		
	PESQ	LSD	SSNR	PESQ	LSD	SSNR
SNR20	3.39	2.30	8.42	**3.40**	**2.27**	**8.44**
SNR15	**3.10**	2.90	6.34	**3.10**	**2.87**	**6.40**
SNR10	**2.80**	**4.08**	**3.86**	2.79	**4.08**	3.85
SNR5	**2.46**	**5.87**	**1.23**	2.46	5.92	1.17
SNR0	**2.10**	**8.23**	**-1.42**	2.08	8.32	-1.50
SNR-5	**1.74**	**11.00**	**-3.86**	1.73	11.14	-3.96
Ave	**2.60**	**5.73**	**2.43**	2.59	5.77	2.40

B. Evaluation of the Three Proposed Strategies for Unseen Noise Environments

As shown in Eqs. (6)-(7), two GV equalization factors were proposed to sharpen the over-smoothed estimated clean speech spectra. In Table V we compare the PESQ results of the DNN baseline and GV equalization using factors, α and β, on the test set at different SNRs of the three unseen noise environments. The performance of GV equalization outperformed the DNN baseline, especially at high SNRs. Using the dimension-independent factor β consistently outperformed that using the dimension-dependent factor α indicates that the same scaling factor could be used for each frequency bin. Nonetheless, the values of the factor α of different bins were fluctuant, especially at low (and high) frequencies. This might lead to unreasonable stretch of the estimated speech spectra.

TABLE V
PESQ RESULTS OF THE DNN BASELINE AND GV EQUALIZATION USING FACTOR α AND β ON THE TEST SET AT DIFFERENT SNRS OF THREE UNSEEN NOISE ENVIRONMENTS.

	Exhibition			Destroyer engine			HF channel		
	DNN	α	β	DNN	α	β	DNN	α	β
SNR20	3.37	3.50	**3.53**	3.51	3.62	**3.67**	3.27	**3.41**	**3.41**
SNR15	3.09	3.20	**3.22**	3.27	3.36	**3.41**	2.95	**3.06**	**3.06**
SNR10	2.78	2.86	**2.87**	3.00	3.08	**3.13**	2.61	2.69	**2.70**
SNR5	2.41	2.47	**2.49**	2.69	2.76	**2.81**	2.28	2.34	**2.35**
SNR0	1.99	2.03	**2.05**	2.35	2.42	**2.46**	1.96	2.00	**2.02**
SNR-5	1.57	1.58	**1.61**	2.00	2.06	**2.11**	1.65	1.70	**1.72**
Ave	2.54	2.61	**2.63**	2.80	2.88	**2.93**	2.45	2.53	**2.54**

TABLE VI
PESQ RESULTS OF THE DNN BASELINE AND USING NAT ON THE TEST SET AT DIFFERENT SNRS OF THREE UNSEEN NOISE ENVIRONMENTS.

	Exhibition		Destroyer engine		HF channel	
	DNN	NAT	DNN	NAT	DNN	NAT
SNR20	3.37	**3.43**	3.51	**3.59**	3.27	**3.27**
SNR15	3.09	**3.15**	3.27	**3.34**	**2.95**	2.93
SNR10	2.78	**2.84**	3.00	**3.06**	2.61	**2.61**
SNR5	2.41	**2.46**	2.69	**2.74**	2.28	**2.31**
SNR0	1.99	**2.03**	2.35	**2.39**	1.96	**2.03**
SNR-5	1.57	**1.60**	2.00	**2.04**	1.65	**1.74**
Ave	2.54	**2.59**	2.80	**2.86**	2.45	**2.48**

Table VI presents the PESQ results of using NAT on the test set at different SNRs of three unseen noise environments. The DNN using NAT outperformed the DNN baseline at almost all conditions, e.g., an average PSEQ improvement of 0.06 in the *Destroy engine* noise.

In Table VII, we compare the PESQ results among the noisy, denoted as (a), the LogMMSE enhanced, denoted as (b), the DNN baseline enhanced, denoted as (c), the dropout DNN enhanced, denoted as (d), the GV equalization DNN enhanced, denoted as (e), the dropout and GV equalization DNN enhanced, denoted as (f) and the jointly dropout, GV equalization and NAT DNN enhanced, denoted as (g), on the test set at different SNRs in three unseen noise environments. The DNNs were trained by 100 hours of training data and 104 noise types, with 3 hidden layers, 2048 units in each hidden layer, and 11-frame acoustic context. Compared with the DNN baseline system where only the basic DNN training procedure is applied, the system improved by dropout training indeed showed better performances, with average PESQ going from 2.45 in column (c) to 2.53 in column (d) for *HF channel* noise, especially at low SNRs, with PESQ going from 1.65 to 1.80 for SNR=-5dB in *HF channel* noise.

Meanwhile, GV equalization also achieved significant improvements over the DNN baseline, with average PESQ going from 2.54 in column (c) to 2.63 in column (e) for *Exhibition* noise, especially at high SNRs, with PESQ going from 3.37 to 3.53 for SNR=20dB in *Exhibition* noise. After jointly improved by GV equalization and dropout, PESQ further increased consistently, with average PESQ going from 2.54 in column (c) to 2.67 in column (f) for *Exhibition* noise, from 2.80 in column (c) to 2.90 in column (f) for *Destroyer engine*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2014.2364452, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE VII
PESQ COMPARISON ON THE TEST SET AT DIFFERENT SNRS OF UNSEEN NOISE ENVIRONMENTS, AMONG: (A) NOISY,(B) LOGMMSE APPROACH, (C) DNN BASELINE, (D) DNN WITH DROPOUT, (E) DNN WITH GV EQUALIZATION, (F) DNN WITH DROPOUT AND GV EQUALIZATION, AND (G) DNN WITH JOINT DROPOUT, GV EQUALIZATION AND NAT.

	Exhibition							Destroyer engine							HF channel						
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(a)	(b)	(c)	(d)	(e)	(f)	(g)
SNR20	2.89	3.24	3.37	3.38	3.53	3.44	**3.59**	2.98	3.51	3.51	3.52	3.67	3.58	**3.76**	2.77	3.37	3.27	3.28	3.41	3.32	**3.47**
SNR15	2.55	2.91	3.09	3.12	3.22	3.18	**3.31**	2.66	3.21	3.27	3.29	3.41	3.36	**3.53**	2.43	3.08	2.95	2.98	3.06	3.02	**3.15**
SNR10	2.21	2.58	2.78	2.82	2.87	2.90	**3.00**	2.34	2.87	3.00	3.03	3.13	3.10	**3.24**	2.10	2.74	2.61	2.66	2.70	2.72	**2.82**
SNR5	1.87	2.15	2.41	2.47	2.49	2.55	**2.63**	2.04	2.49	2.69	2.73	2.81	2.80	**2.91**	1.78	2.33	2.28	2.37	2.35	2.43	**2.52**
SNR0	1.56	1.69	1.99	2.08	2.05	2.18	**2.24**	1.75	2.07	2.35	2.39	2.46	2.46	**2.55**	1.51	1.84	1.96	2.08	2.02	2.15	**2.24**
SNR-5	1.28	1.23	1.57	1.65	1.61	1.75	**1.80**	1.52	1.57	2.00	2.04	2.11	2.10	**2.17**	1.31	1.33	1.65	1.80	1.72	1.84	**1.92**
Ave	2.06	2.30	2.54	2.59	2.63	2.67	**2.76**	2.22	2.62	2.80	2.83	2.93	2.90	**3.03**	1.99	2.45	2.45	2.53	2.54	2.58	**2.69**

noise, and from 2.45 in column (c) to 2.58 in column (f) for *HF channel* noise.

By incorporating NAT on top of dropout and GV equalization the best average PESQ results were achieved in columns (g) at all three unseen noise types. It is clear that the three techniques were complementary by a PESQ comparison from columns (c) to (g). Furthermore, the best DNN system significantly outperformed the LogMMSE method (achieving only 2.30, 2.62 and 2.45 of average PESQ in columns (b) for all three noise types, respectively) at different SNR levels of all noise conditions, especially at low SNRs for the noise type with many non-stationary components, e.g., PESQ going from 1.69 in column (b) to 2.24 in column (g) under *Exhibition* noise at SNR=0dB.

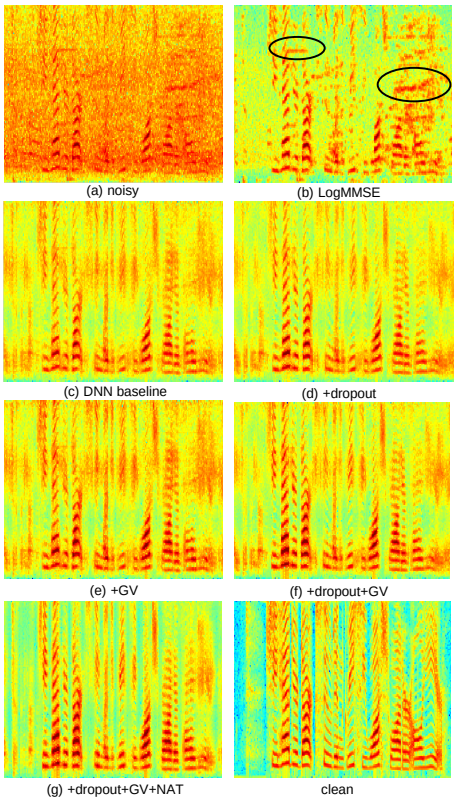


Fig. 8. Spectrograms of an utterance tested with *Exhibition* noise at SNR = 5dB. (a) noisy speech (PESQ=1.42), (b) LogMMSE (PESQ=1.83), (c) DNN baseline (PESQ=1.87), (d) improved by dropout (PESQ=2.06), (e) improved by GV equalization (PESQ=2.00), (f) improved by dropout and GV (PESQ=2.13), (g) jointly improved by dropout, NAT and GV equalization (PESQ=2.25), and the clean speech (PESQ=4.5).

The enhanced spectrograms from one noisy speech utterance corrupted by *Exhibition* noise at SNR=5dB using different techniques were shown in Fig. 8. First, the LogMMSE method played a limited role in reducing the non-stationary noise components and there was still a lot of scatter noise in the enhanced spectrogram, as shown in the two circled regions in Fig. 8(b). Second, although the non-stationary noise components in the noisy spectra shown in Fig. 8(a) disappeared after processing by DNN shown in Fig. 8(c), some residual noise still existed, as the *Exhibition* noise was unseen in the training set. By a comparison from Fig. 8(c) to Fig. 8(g), we could observe that dropout and NAT techniques could reduce this relatively stationary residue noise in Fig. 8(c). while the enhanced formant spectra could be brightened using GV equalization. The final spectrogram enhanced by DNN in Fig. 8(g) obviously seemed more noiseless than that using LogMMSE in Fig. 8(b), with a reference clean speech spectrogram at the bottom-right corner of Fig. 8.

C. Suppression against Non-stationary Noise

It was of a great interest to examine the effect of DNN against non-stationary noise, which is quite common in real-world noisy speech. Fig. 9 shows an utterance example corrupted by *Machine gun* noise at SNR=-5dB. It was known to be difficult for almost all of the conventional techniques to track the sudden increases of noise power, or they are overestimating the noise energy resulting in speech distortion [41]. The LogMMSE method did not work under this burst noise at all, achieving PESQ=1.86 which is almost the same as the PESQ value of 1.85 for noisy speech. Even using the training data with only four noise types, the trained DNN still had a strong suppression ability against non-stationary noises, achieving PESQ=2.14. Finally the DNN trained with 104 noise types obtained a good effect in listening quality and with the best PESQ value of 2.78. This demonstrated that using a DNN model, with an adequate acoustic context (both in time and in frequency) and trained with a large coverage of noise types, can well deal with the unseen noise type, especially for the non-stationary noise components.

Fig. 10 shows an utterance example corrupted in succession by different noise types at several speech segments. These noise types were *Exhibition*, *Buccaneer2*, *F16*, *Leopard*, and *Destroyer engine*. The DNN-enhanced spectrogram shown in Fig. 10(a) successfully removed most of the noises while the LogMMSE-enhanced spectrogram shown in Fig. 10(b) failed to remove most of them and even led to a worse PESQ than

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2014.2364452, IEEE/ACM Transactions on Audio, Speech, and Language Processing

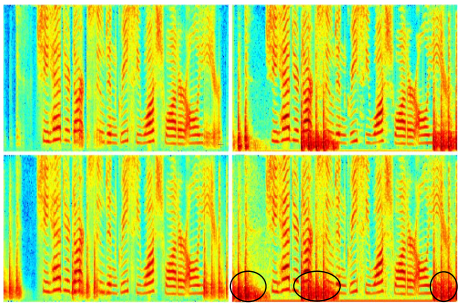


Fig. 9. Spectrograms of an utterance tested on *Machine gun* noise at SNR = -5dB: with 104-noise DNN enhanced (upper left, PESQ=2.78), LogMMSE enhanced (upper right, PESQ=1.86), 4-noise DNN enhanced (bottom left, PESQ=2.14), and noisy speech (bottom right, PESQ=1.85).

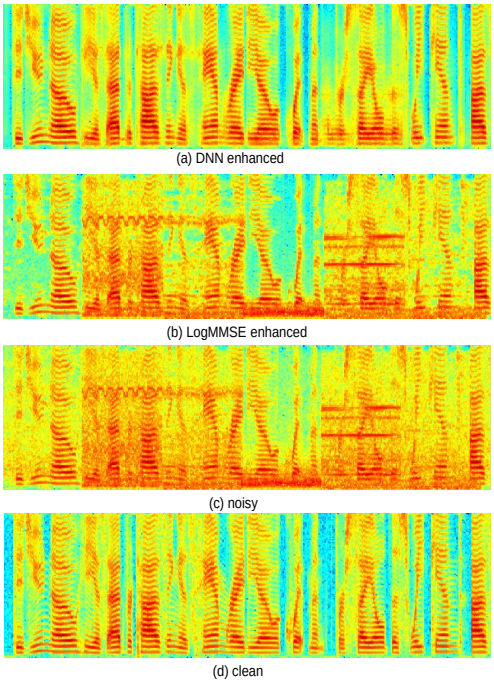


Fig. 10. Spectrograms of an utterance corrupted in succession by different noise types tested on changing noise environments at SNR = 5dB: (a) DNN enhanced (PESQ=2.99), (b) the LogMMSE enhanced (PESQ=1.46), (c) noisy (PESQ=2.05), and (d) clean speech (PESQ=4.50).

the noisy speech (PESQ going down from 2.05 to 1.46). This was reasonable as the LogMMSE method predicted the noise in a recursive averaging mode according to previous frames and it was hard to track the potentially dramatic changes in non-stationary noises. However, the DNN model processed the noisy spectrum in a frame-by-frame manner, and the relationship between the clean speech and noise had been learned off-line. As handling non-stationary noises is still an open research problem in speech enhancement, our study gives a possible research direction to solve it.

D. Overall Evaluation

The overall evaluation results on the test set with the whole 15 unseen noise types[1], among the LogMMSE, the DNN baseline with 100 hours of data, the improved DNN with 100 hours of data and the improved DNN with 625 hours of data, are listed in Table VIII. All DNN configurations were fixed at $L = 3$ hidden layers, 2048 units at each hidden layer, and 11-frame input. The DNN baseline could be improved effectively using the three proposed techniques discussed in Section III. It is interesting to note that the best average PSEQ of 3.15 was achieved with 625 hours of stereo training data. A larger training set was shown to be slightly better than the situation with a smaller training set of 100 hours (achieving an average PESQ of 3.12). Moreover, we can see that the absolute PESQ gained between our best DNN system and LogMMSE system (0.37) was even comparable to that between LogMMSE system and unprocessed noisy speech system (0.38), which was believed to be a significant improvement. Finally, by using more noise types and the three proposed techniques, the PESQ improvements of the proposed DNN approach over LogMMSE under unseen noise types in Table VIII are also comparable to that under matched noise types reported in [32]. Meanwhile, the STOI results to represent the intelligibility of the enhanced speech were also presented in Table IX. LogMMSE is slightly better than the noisy with an average STOI improvement from 0.81 to 0.82. The DNN baseline trained with 100 hours got 0.86 STOI score on average. The proposed strategies could further improve the performance. After trained with 625 hours data, the STOI was improved to 0.88, especially at low SNRs. As for the intelligent of the speech, we may care more about the low SNR conditions. Although there is a little performance degradation at SNR=20dB, an absolute 0.13 STOI improvement compared with the LogMMSE method was obtained at SNR=-5dB. More results and demos can be found at this website[3].

TABLE VIII
AVERAGE PESQ RESULTS AMONG THE LOGMMSE, THE DNN BASELINE WITH 100 HOURS DATA, THE IMPROVED DNN WITH 100 HOURS DATA AND THE IMPROVED DNN WITH 625 HOURS DATA ON THE TEST SET AT DIFFERENT SNRS ACROSS THE WHOLE 15 UNSEEN NOISE TYPES.

	Noisy	LogMMSE	100h-baseline	100h-impr	625h-impr
SNR20	3.21	3.60	3.62	3.77	**3.80**
SNR15	2.89	3.33	3.39	3.58	**3.60**
SNR10	2.57	3.02	3.13	3.33	**3.36**
SNR5	2.24	2.66	2.85	3.05	**3.08**
SNR0	1.91	2.25	2.52	2.71	**2.74**
SNR-5	1.61	1.80	2.16	**2.31**	**2.31**
Ave	2.40	2.78	2.94	3.12	**3.15**

TABLE IX
AVERAGE STOI RESULTS AMONG THE LOGMMSE, THE DNN BASELINE WITH 100 HOURS DATA, THE IMPROVED DNN WITH 100 HOURS DATA AND THE IMPROVED DNN WITH 625 HOURS DATA ON THE TEST SET AT DIFFERENT SNRS ACROSS THE WHOLE 15 UNSEEN NOISE TYPES.

	Noisy	LogMMSE	100h-baseline	100h-impr	625h-impr
SNR20	**0.97**	**0.97**	0.96	0.96	0.96
SNR15	0.93	0.94	**0.95**	**0.95**	**0.95**
SNR10	0.88	0.89	0.92	0.92	**0.93**
SNR5	0.80	0.81	0.87	0.88	**0.89**
SNR0	0.70	0.70	0.79	0.81	**0.82**
SNR-5	0.60	0.58	0.70	**0.71**	**0.71**
Ave	0.81	0.82	0.86	0.87	**0.88**

[3]http://home.ustc.edu.cn/~xuyong62/demo/SE_DNN_taslp.html.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2014.2364452, IEEE/ACM Transactions on Audio, Speech, and Language Processing

E. Evaluation for Real-world Noise Environments

Table X shows the informal subjective preference evaluation comparison between DNN enhanced and LogMMSE enhanced speech for 32 real-world noisy utterances (22 spoken in English, and others spoken in other languages), which were collected from some movies, lectures, or recorded directly by the authors. They were assigned to ten subjects (five Chinese males and five Chinese females.) for listening preference choices. An average of 78% of the subjects preferred DNN enhanced speech. For testing on English which is the same language as in the TIMIT utterances used for training, the preference score was 81%, higher than the score of 75% for those utterances in different languages. Although DNNs did well in cross-language testing, more research is needed to bridge this performance gap.

TABLE X
SUBJECTIVE PREFERENCE EVALUATION COMPARISON BETWEEN THE DNN ENHANCED AND LOGMMSE ENHANCED SPEECH OF 32 REAL-WORLD NOISY UTTERANCES IN ENGLISH OR OTHER LANGUAGES.

	English	Others	Ave
DNN	81%	75%	78%
LogMMSE	19%	25%	22%

Finally to illustrate the speech quality obtained with real-world noisy utterances we present testing results for an utterance extracted from the famous movie ***Forrest Gump*** and spoken by the well-known actor ***Tom Hanks*** playing the title role. In Fig. 11 the spectrograms corresponding to the best DNN model, the LogMMSE and the noisy speech are shown. It was observed that the DNN model could still well handle the particular noisy condition. Compared to the LogMMSE-enhanced speech shown in the middle panel, the DNN-enhanced speech (shown in the left panel) was seen to suppress non-stationary noise more and resulted in less residual noise.

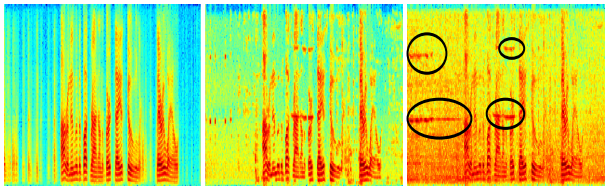


Fig. 11. Spectrograms of a noisy utterance extracted from the movie ***Forrest Gump*** with: improved DNN (left), LogMMSE (middle) and noisy speech (right).

V. CONCLUSION

In this paper, a DNN-based framework for speech enhancement is proposed. Among the various DNN configurations, a large training set is crucial to learn the rich structure of the mapping function between noisy and clean speech features. It was found that the application of more acoustic context information improves the system performance and makes the enhanced speech less discontinuous. Moreover, multi-condition training with many kinds of noise types can achieve a good generalization capability to unseen noise environments. By doing so, the proposed DNN framework is also powerful to cope with non-stationary noises in real-world environments. An over-smoothing problem in speech quality was found in the MMSE-optimized DNNs and one proposed post-processing technique, called GV equalization, was effective in brightening the formant spectra of the enhanced speech signals. Two improved training techniques were further adopted to reduce the residual noise and increase the performance. Compared with the LogMMSE method, significant improvements were achieved across different unseen noise conditions. Another interesting observation was that the proposed DNN-based speech enhancement system is quite effective for dealing with real-world noisy speech in different languages and across different recording conditions not observed during DNN training.

It should be noted that only the TIMIT corpus was used to construct the clean speech training set in the current study. Such a small amount of data cannot be expected to attain a good coverage of different acoustic conditions, such as speaker and language variabilities. In future studies, we would increase the speech diversity by first incorporating clean speech data from a rich collection of materials covering more languages and speakers. Second, there are many factors in designing the training set. We would utilize principles in experimental design [54, 55] for multi-factor analysis to alleviate the requirement of a huge amount of training data and still maintain a good generalization capability of the DNN model. Third, some other features, such as Gammatone filterbank power spectra [50], Multi-resolution cochleagram feature [56], will be adopted as in [50] to enrich the input information to DNNs. Finally, a dynamic noise adaptation scheme will also be investigated for the purpose of improving tracking of non-stationary noises.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC press, 2013.
[2] J. Benesty, S. Makino, and J. D. Chen, *Speech Enhancement*, Springer, 2005.
[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustic, Speech and Signal Processing*, Vol. 27, No. 2, pp. 113-120, 1979.
[4] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," in *Proc. IEEE*, Vol. 67, No. 12, pp. 1586-1604, 1979.
[5] P. Scalart, "Speech enhancement based on a priori signal to noise estimation," in *Proc. ICASSP*, pp. 629-632, 1996.
[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 32, No.6, pp. 1109-1121, 1984.
[7] Y. Ephraim and D. Malah, "Speech enhancement using minimu mean square log spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 33, No. 2, pp. 443-445, 1985.
[8] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, Vol. 81, No. 11, pp. 2403-2418, 2001.
[9] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging,"

IEEE Trans. on Speech and Audio Processing, Vol. 11, No. 5, pp. 466-475, 2003.

[10] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp. 345-349, 1994.

[11] A. Hussain, M. Chetouani, S. Squartini, A. Bastari, and F. Piazza, "Nonlinear speech enhancement: an overview," in *Progress in Nonlinear Speech Processing*, Springer, pp. 217-248, 2007.

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

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

[14] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," in *Handbook of Neural Networks for Speech Processing*, Edited by Shigeru Katagiri, Artech House, Boston, 1998.

[15] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Trans. on Speech and Audio Processing*, Vol. 11, No. 3, pp. 184-192, 2003.

[16] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, Vol. 2, No. 1, pp. 1-127, 2009.

[17] D. Erhan, A. Courville, Y. Bengio and P. Vincent, "Why does unsupervised pre-training help deep learning?" *The Journal of Machine Learning Research*, No. 11, pp. 625-660, 2010.

[18] G. E. Hinton, S. Osindero and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, Vol. 18, No. 7, pp. 1527-1554, 2006.

[19] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, Vol. 313, No. 5786, pp. 504-507, 2006.

[20] L.-H. Chen, Z.-H. Ling, L.-J. Liu and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Trans. on Audio, Speech and Language Processing*, Vol. 22, No. 12, pp. 1859-1872, 2014.

[21] Z.-H. Ling, L. Deng and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 21, No. 10, pp. 2129-2139, 2013.

[22] B.-Y. Xia and C.-C. Bao, "Speech enhancement with weighted denoising Auto-Encoder," *Proc. Interspeech*, pp. 3444-3448, 2013.

[23] X.-G. Lu and Y. Tsao and S. Matsuda and C. Hori, "Speech enhancement based on deep denoising Auto-Encoder," *Proc. Interspeech*, pp. 436-440, 2013.

[24] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," *Proc. Interspeech*, pp. 22-25, 2012.

[25] M. Wollmer, Z. Zhang, F. Weninger, B. Schuller and G. Rigoll, "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise," *Proc. ICASSP*, pp. 6822-6826, 2013.

[26] H. Christensen, J. Barker, N. Ma and P. D. Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," *Proc. Interspeech*, pp. 1918-1921, 2010.

[27] Y. X. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 21, No. 7, pp. 1381-1390, 2013.

[28] E. W. Healy, S. E. Yoho, Y. X. Wang and D. L. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, Vol. 134, No. 4, pp. 3029-3038, 2013.

[29] D. L. Wang and G. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Hoboken, NJ, USA: Wiley-IEEE Press, 2006.

[30] A. Narayanan, D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," *Proc. ICASSP*, pp. 7092-7096, 2013.

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

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

[33] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, Vol. 48, No. 2, pp. 220-231, 2006.

[34] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," http://arxiv.org/abs/1207.0580, 2012.

[35] M. Seltzer, D. Yu and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," *Proc. ICASSP*, pp. 7398-7402, 2013.

[36] A. Narayanan and D. L. Wang, D, "Joint noise adaptive training for robust automatic speech recognition," *Proc. ICASSP*, pp. 2523-2527, 2014.

[37] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," *Proc. ICASSP*, pp. 7942-7946, 2013.

[38] A. Varga and H. J. M. 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.

[39] G. Shi, M. M. Shanechi and P. Aarabi, "On the importance of phase in human speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 14, No. 5, pp. 1867-1874, 2006.

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

[41] S. Rangachari, P. C. Loizou and Y. Hu, "A noise estimation algorithm with rapid adaptation for highly nonstationary environments," *Proc. ICASSP*, pp. 305-308, 2004.

[42] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A. R. Mohamed and G. E. Hinton, "Binary coding of speech spectrograms using a deep auto-encoder," *Proc. Interspeech*, pp. 1692-1695, 2010.

[43] T. Toda, A. W. Black and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," *Proc. ICASSP*, pp. 9-12, 2005.

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

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

[46] ITU-T, Recommendation 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," *International Telecommunication Union-Telecommunication Standardisation Sector*, 2001.

[47] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "An algorithm for intelligibility prediction of timefrequency weighted noisy speech," *IEEE Trans. on Speech and Audio Processing*, Vol. 19, No. 7, pp. 2125-2136, 2011.

[48] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech*, Prentice Hall, 2010.

[49] G. Hu, 100 nonspeech environmental sounds, 2004. http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html.

[50] Y. Wang, A. Narayanan, D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. on Speech and Audio Processing*, Vol. 22, No. 12, pp. 1849-1858, 2014.

[51] H. Larochelle, Y. Bengio, J. Louradour and P. Lamblin, "Exploring strategies for training deep neural networks," *The Journal of Machine Learning Research*, No. 10, pp. 1-40, 2009.

[52] S. Srinivasan, *Knowledge-Based Speech Enhancement*, Ph.D. thesis, KTH, 2005.

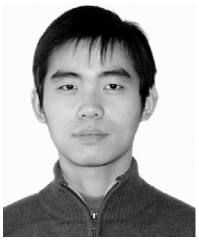
[53] A. D. L. Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 13, No. 3, pp. 355-366, 2005.

[54] M. J. Anderson and P. J. Whitcomb, *Design of Experiments*, John Wiley & Sons, Inc., 1974.

[55] R. L. Plackett and J. P. Burman, "The design of optimum multifactorial experiments," *Biometrica*, Vol. 33, No. 4, pp. 305-325, 1946.

[56] X. L. Zhang and D. L. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," *Proc. Interspeech*, pp. 1534-1538, 2014.

Yong Xu received the B.S. degree in communication engineering from Anhui University in 2010. He is currently a Ph.D. candidate of University of Science and Technology of China (USTC). From Jul. 2012 to Dec. 2012, he was an intern at iFlytek. From Sept. 2014 to April 2015, he is a visiting student at Georgia Institute of Technology, USA. His current research interests include deep learning for speech enhancement and noise robust speech recognition.



Jun Du received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC) in 2004 and 2009, respectively. From 2004 to 2009, he was with the iFlytek Speech Lab of USTC. During the above period, he worked as an Intern twice for 9 months at Microsoft Research Asia (MSRA), Beijing. In 2007, he also worked as a Research Assistant for 6 months at the Department of Computer Science, The University of Hong Kong. From July 2009 to June 2010, he worked at iFlytek Research on speech recognition. From July 2010 to January 2013, he joined MSRA as an Associate Researcher, working on handwriting recognition, OCR, and speech recognition. Since February 2013, Dr. Du worked at National Engineering Laboratory for Speech and Language Information Processing (NEL-SLIP) of USTC.

Li-Rong Dai was born in China in 1962. He received the B.S. degree in electrical engineering from Xidian University, Xian, China, in 1983 and the M.S. degree from Hefei University of Technology, Hefei, China, in 1986, and the Ph.D. degree in signal and information processing from the University of Science and Technology of China (USTC), Hefei, in 1997. He joined University of Science and Technology of China in 1993. He is currently a Professor of the School of Information Science and Technology, USTC. His current research interests include speech synthesis, speaker and language recognition, speech recognition, digital signal processing, voice search technology, machine learning, and pattern recognition. He has published more than 50 papers in these areas.

Chin-Hui Lee is a professor at School of Electrical and Computer Engineering, Georgia Institute of Technology. Before joining academia in 2001, he had 20 years of industrial experience ending in Bell Laboratories, Murray Hill, New Jersey, as a Distinguished Member of Technical Staff and Director of the Dialogue Systems Research Department. Dr. Lee is a Fellow of the IEEE and a Fellow of ISCA. He has published over 400 papers and 30 patents, and was highly cited for his original contributions with an h-index of 66. He received numerous awards, including the Bell Labs President's Gold Award in 1998. He won the SPS's 2006 Technical Achievement Award for "Exceptional Contributions to the Field of Automatic Speech Recognition". In 2012 he was invited by ICASSP to give a plenary talk on the future of speech recognition. In the same year he was awarded the ISCA Medal in scientific achievement for pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition.