A Reverberation-Time-Aware Approach to Speech Dereverberation Based on Deep Neural Networks

Bo Wu, Kehuang Li, Minglei Yang, and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*—A reverberation-time-aware deep-neural-network (DNN)-based speech dereverberation framework is proposed to handle a wide range of reverberation times. There are three key steps in designing a robust system. First, in contrast to sigmoid activation and min–max normalization in state-of-the-art algorithms, a linear activation function at the output layer and global mean-variance normalization of target features are adopted to learn the complicated nonlinear mapping function from reverberant to anechoic speech and to improve the restoration of the low-frequency and intermediate-frequency contents. Next, two key design parameters, namely, frame shift size in speech framing and acoustic context window size at the DNN input, are investigated to show that RT60-dependent parameters are needed in the DNN training stage in order to optimize the system performance in diverse reverberant environments. Finally, the reverberation time is estimated to select the proper frame shift and context window sizes for feature extraction before feeding the log-power spectrum features to the trained DNNs for speech dereverberation. Our experimental results indicate that the proposed framework outperforms the conventional DNNs without taking the reverberation time into account, while achieving a performance only slightly worse than the oracle cases with known reverberation times even for extremely weak and severe reverberant conditions. It also generalizes well to unseen room sizes, loudspeaker and microphone positions, and recorded room impulse responses.**

***Index Terms*—Acoustic context, deep neural networks (DNNs), frame shift, linear output layer, mean-variance normalization, reverberation-time-aware (RTA), speech dereverberation.**

I. Introduction

When a microphone is placed at a distance from a talker in an enclosed space in hands-free, eyes-busy speech applications, the received signal will be a collection of many delayed and attenuated copies of the original speech signals, caused by the reflections from walls, ceilings, and floors [1]. As a result, reverberation often seriously degrades speech quality and intelligibility. Such deteriorations can cause decreased performances for automatic speech recognition, hearing aids and source localization. Thus, an effective dereverberation solution will benefit many speech applications.

Many dereverberation techniques have been proposed in the past (e.g., [2]–[6]). One direct way is to estimate an inverse filter of the room impulse response (RIR) [7] to deconvolve the reverberant signal. However, a minimum phase assumption is often needed, which is almost never satisfied in practice [7]. The RIR can also be varying in time and hard to estimate [1]. The work presented in [2] estimated a fixed length of an inverse filter of RIR by maximizing the kurtosis of the linear prediction (LP) residual for the reduction of early reverberation, without taking into account the impact of RIR in distinct reverberant environments on the system performance. The inverse filtering is only effective in a short reverberation time (RT60) [8] range, from 0.2 to 0.4 s. Mosayyebpour *et al.* [4] presented an iterative method to blindly determine the filter length according to the reverberant condition, which could be used in highly reverberant rooms. Nevertheless, the stopping criterion was empirically chosen. Kinoshita *et al.* [3] estimated the late reverberations using long-term multi-step linear prediction, and then reduced the late reverberation effect by employing spectral subtraction.

Some studies attempted to separate speech and reverberation via homomorphic transformation [9], [10]. Nevertheless, they are not very effective when the human auditory system is the target. Other methods dealt with dereverberation by exploiting the essential properties of speech such as harmonic filtering [11]. The dereverberation filter is only estimated from voiced speech segments, therefore achieving a poor dereverberation performance for unvoiced speech segments.

Recently, due to their strong regression capabilities, deep neural networks (DNNs) [12], [13] have also been utilized in speech enhancement [14], [15], source separation[16], [17] and bandwidth expansion [18], [19]. Han *et al.* [5], [20] also proposed to dereverberate speech using DNNs, to learn a spectral mapping from reverberant to anechoic speech. Although the results reflect the state-of-the-art performances, they represented the DNN prediction of log-spectral magnitude into an unit range and normalized the target features into the same range, preventing a good dereverberation performance, especially at low RT60s. Moreover, their system is environmentally insensitive, not being to realize its full potential.

In this study, we utilize an improved DNN dereverberation system we proposed recently [21] by adopting a linear output layer and globally normalizing the target features into zero mean and unit variance, and then investigate the effects of frame shift and acoustic context sizes on the dereverberated speech quality using DNNs at different RT60s. We show that on the one hand low frame shifts can not obtain good performances in

Manuscript received April 29, 2016; revised August 8, 2016 and October 20, 2016; accepted October 20, 2016. Date of publication October 31, 2016; date of current version November 28, 2016. This work was supported in part by the National Natural Science Foundation of China (61571344). The work of B.Wu was supported by a grant from the China Scholarship Council. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. DeLiang Wang.

B. Wu and M. Yang are with the National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710126, China (e-mail: rambowu11@gmail.com; mlyang@xidian.edu.cn).

K. Li and C.-H. Lee are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: kehlekernel@gmail.com; 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/TASLP.2016.2623559

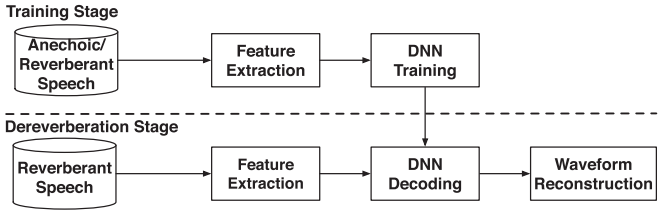


Fig. 1. A DNN-based speech dereverberation system.

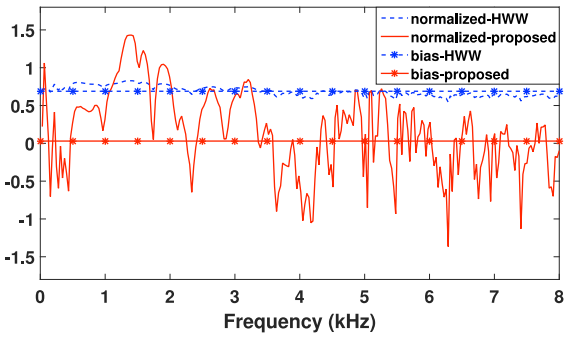


Fig. 2. Plots of normalized log-power spectra of a target frame with HWW-DNN (in dash curve and line) and the proposed DNN (in solid curve and line). "normalized" and "bias" denote each spectrum and its mean, respectively.

a strong reverberant condition, even at the price of increased computational complexities. We next demonstrate that on the other hand a large number of speech frames covering a large acoustic context commonly used at the DNN input often degrades the quality of dereverberated speech at a low RT60. Based on these observations, a frame-shift-aware DNN (FSA-DNN) and an acoustic-context-aware DNN (ACA-DNN) are designed by adopting RT60-dependent frame shift and acoustic context sizes as key DNN design parameters. We further explore a reverberation-time-aware DNN (RTA-DNN) that outperforms separate systems by considering both effects.

The rest of the paper is organized as follows. We first describe the proposed RTA-DNN dereverberation system in Section II. Motivations for exploring the frame shift and acoustic context parameters in reverberant situation are given in Section III. Experimental results are next provided and analyzed in Section IV. The generalization capabilities of the proposed DNN models are illustrated in Section V. Finally we summarize our findings in Section VI.

II. DNN-Based Speech Dereverberation Model

A block diagram of the DNN-based speech dereverberation system is illustrated in Fig. 1. In the training stage, a regression DNN [14] is trained by a set of multi-condition data, consisting of pairs of reverberant and anechoic speech represented by log-power spectra (LPS). In the dereverberation stage, the well-trained DNN is fed with the LPS features of input speech to generate the corresponding enhanced LPS features. The required phase is directly extracted from the reverberant speech, because human ears are considered to be not sensitive to such information [22]. Finally the dereverberated waveform is reconstructed from the estimated spectral magnitude and the reverberant speech phase with an overlap-add method [23].

A. Output Layer Activation and Target Feature Normalization

1) Sigmoid Activation and Min–Max Normalization: In [5], Han *et al.* proposed to learn the log-spectral mapping function from reverberant to anechoic speech using a nonlinear DNN-based regression model, and generated the state-of-the-art dereverberation performances. They represented the DNN output of log-spectral magnitude into an unit range of [0, 1] by using a sigmoid activation function and normalized the target (anechoic) features into the same range. A minimum mean squared error (MMSE) objective function between the DNN output and target features can be illustrated as follows:

$$\min E = \frac{1}{ND}\sum_{n=1}^{N}\sum_{d=1}^{D}||\hat{X}_d^n - X_{d,\mathrm{norm}}^n||_2^2 \tag{1}$$

where E denotes mean squared error. $\hat{X}_d^n$ and $X_{d,\mathrm{norm}}^n$ represent the d-th DNN output and normalized target feature at frame index n, with D and N denoting the size of feature vector and mini-batch, respectively.

$$\hat{X}_d^n = \frac{1}{1+e^{-(\sum_{m=1}^{M} W_{\mathrm{md}} S_m^n + b_d)}} \tag{2}$$

$$X_{d,\mathrm{norm}}^n = \frac{X_d^n - X_{\min}}{X_{\max} - X_{\min}} \tag{3}$$

where W_{md} and b_d represent the weights and biases between the last hidden layer and output layer, respectively, with M denoting the last hidden layer size. S_m^n is the output at the m-neuron of the last hidden layer at frame index n. $X_{\max}$ and $X_{\min}$ denote the maximum and minimum values of the spectral feature among all target utterances, respectively. There is no obvious difference between using dimension dependent min/max value vector and using the min/max value calculated among all dimensions in Eq. (3) in our experiments.

The dash curve in Fig. 2 illustrates the normalized log-power spectrum of a target frame, $X_{d,\mathrm{norm}}^n$, using Eq. (3). We refer to the DNN obtained here as HWW-DNN. It can be seen that the dynamic range here is small, resulting in blurred harmonics and thus preventing an accurate restoration of the estimated anechoic spectrogram. In addition, the DNN also needs to learn the non-zero bias of the target feature vectors. This scheme seriously degrades the dereverberation performances, especially when the training set size is small.

2) Linear Activation and Mean-Variance Normalization: To deal with the abovementioned drawbacks of the target feature mapping scheme discussed in Section II-A1, we propose a linear activation function at the output layer of the DNN and to globally normalize the target features over all the target utterances into zero mean and unit variance [21]. This is the popular mean variance normalization (MVN) strategy [24] commonly used in the speech recognition community. The DNN output and

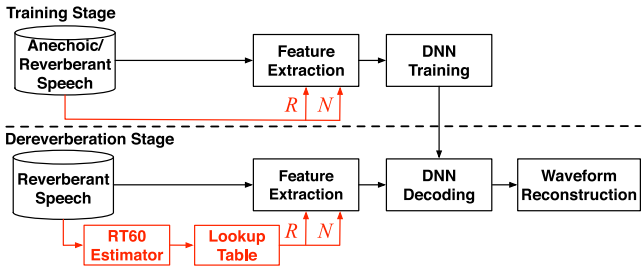


Fig. 3. A block diagram of the proposed RTA-DNN dereverberation system.

normalized target features are:

$$\hat{X_d^n} = \sum_{m=1}^{M} W_{\mathrm{md}} S_m^n + b_d \tag{4}$$

$$X_{d,\mathrm{norm}}^n = \frac{X_d^n - \mu_d}{\sigma_d} \tag{5}$$

where μ_d and σ_d represent the global mean and variance of the target feature at frequency bin d over all target utterances.

As shown in the solid curve in Fig. 2, the proposed normalized log-power spectrum of the target frame clearly retains spectral details and observable harmonics. Therefore, it can boost the restoration of the estimated anechoic spectrogram, especially at low and intermediate frequencies, which will be shown in subsequent experiments later. Furthermore, the bias of the proposed normalized feature is always close to zero, which will release a burden of DNN training and make the proposed system more effective even with less training samples, to be illustrated later in Section IV-B4 .

B. Reverberation-Time-Aware DNN (RTA-DNN)

In signal processing of reverberant speech, different reverberant environments will result in distinct superpositions in the time domain and inter-frame correlations, which is often neglected by some dereverberation algorithms (e.g., [2]–[5]). For the purpose of improving the system performance and enhancing system robustness, we propose an environment-aware approach to take advantage of the characteristics of the superpositions and frame-wise temporal correlations in distinct reverberant situations. An RTA-DNN system is thus designed by adopting two RT60-dependent parameters, namely frame shift size (R) in speech framing and acoustic context size (N) at the DNN input for feature extraction before feeding the log-power spectrum features to the trained DNNs.

A block diagram of the proposed RTA-DNN system is illustrated in Fig. 3 which is an improved version over our proposed DNN system illustrated in Fig. 1, by integrating the two key design parameters, R and N, into training and dereverberation. In the training and dereverberation stages, R and N depend on the utterance-level RT60, while an RT60 estimator followed by a lookup table, which determines R and N according to the estimated RT60, is required in the dereverberation stage. A detailed description of how RT60 affects these two key parameters will be presented later in Sections IV-B1 and IV-B2 and the lookup table will be provided for experiments in Sections IV-B3 and IV-B4.

III. Key Parameters in DNN Dereverberation

A. Frame Shift Size in Speech Framing

Most dereverberation algorithms use the short-time Fourier transform (STFT), $X_n(e^{j\omega})$ [25], which is sampled in both time and frequency dimensions, to obtain a discrete time-frequency representation of speech. Eq. (6) samples $X_n(e^{j\omega})$ at a time rate of (i.e., frame shift) R,

$$X_{rR}(e^{j\omega}) = \sum_{m=-\infty}^{\infty} w(rR-m)x(m)e^{-j\omega m} \tag{6}$$

where r denotes the frame index, $w(rR-m)$ represents the analysis window. Note that, the frame shift R is called the time sampling rate of $X_n(e^{j\omega})$ in [26], different from the time sampling rate T_s of the signal, $x(n)$.

In conventional time sampling of STFT, the frame shift is chosen to avoid an aliased representation of $X_n(e^{j\omega})$ from which $x(n)$ can be exactly recovered [27]. The frame shift is typically fixed to half of the frame length [28] for practical consideration in most dereverberation algorithms.

However, in a reverberant environment, the conventional method is ineffective [2], [3], [5], [29] because it uses a fixed temporal resolution, without considering the mixing conditions of neighboring reverberant frames in different reverberant environments. For weaker reverberation, it takes less time for the reflected sounds to reach back to the microphone [30], resulting in an intensive superposition in the time domain. A dense frame rate, not needed for strong reverberation, is now required to provide a high temporal resolution.

B. Acoustic Context Window Size at DNN Input

In a recently-proposed DNN-based enhancement framework [14] it was demonstrated that using more acoustic context information could improve the continuity of enhanced speech. Nevertheless, in reverberant conditions, the inter-frame correlation depends on RT60 as shown below. The source speech, $x(n)$, and the corresponding received signal at a microphone, $y(n)$, in a reverberant environment can be related by

$$y(n) = x(n) * h(n) \tag{7}$$

where $h(n)$ is the RIR and $*$ denotes convolution. The autocorrelation of $y(n)$ is given as follows [31],

$$\phi_{\mathrm{yy}}(k) = \phi_{\mathrm{xx}}(k) * R_h(k) \tag{8}$$

where $\phi_{\mathrm{yy}}(k)$ and $\phi_{\mathrm{xx}}(k)$ represent the autocorrelation of $y(n)$ and $x(n)$, respectively. And $R_h(k) \triangleq h(k) * h(-k)$ is called the deterministic autocorrelation function of $h(k)$ [32]. Because the length of RIR is proportional to RT60 [33], the lower RT60 is, the shorter RIR is, resulting in a smaller length of $R_h(k)$. Consequently, at a lower RT60, the length of $\phi_{\mathrm{yy}}(k)$ will decrease, because it depends on the length of $R_h(k)$, and the temporal correlation of the consecutive reverberant frames will become weaker. Thus using more context will introduce uncorrelated frames and unnecessary burden in DNN learning [15]. Therefore, an approach exploiting the continuity of the speech

spectra at different RT60s, should be more appropriate than the conventional methods.

IV. Experiments and Result Analysis

A. DNN-Based Speech Dereverberation Models

First, we compare our proposed DNN dereverberation system with HWW-DNN in [5] (without post-processing [34]), which reflects the state-of-the-art performances.

The experiments were conducted in a simulated room of dimension 6 by 4 by 3 meters (length by width by height). The positions of the loudspeaker and the microphone were at (2, 3, 1.5) and (4, 1, 2) meters, respectively. Ten RIRs were simulated using an improved image-source method (ISM) [35] with RT60 ranging from 0.1 to 1.0 s, with an increment of 0.1 s. To learn a high-quality DNN model, all 4620 training utterances from the TIMIT set [36] were convolved with the generated RIRs to build a large training set, resulting in about 40 hours of reverberant speech. Moreover, in order to evaluate the performances of the DNN models at a small training size, a 4-hour training set, that is a subset of the 40-hour training data, was established. To test DNN's generalization capability in mismatch conditions, RIRs with RT60 from 0.1 to 1.0 s, with the increment of 0.05 s (rather than 0.1 s) were convolved with 100 randomly selected utterances from the TIMIT test set to construct the test set. This resulted in a collection of 19×100 reverberant utterances.

For signal analysis, speech was sampled at 16 kHz. The frame length was 32 ms and the frame shift was set to 16 ms according to the conventional frame rate strategy (i.e., frame shift = half of frame length). A 512-point DFT of each overlapping windowed frame is computed. Then 257-dimension log-power spectra feature vectors [23] were used to train DNNs. In addition, perceptual evaluation of speech quality (PESQ) [37], frequency-weighted segmental signal-to-noise ratio (fwSegSNR) [38], [39] and short-time objective intelligibility (STOI) [40] were used to evaluate the dereverberation results.

Kaldi [24] was used to train DNNs, with 3 hidden layers, 2048 nodes for each layer, and 7 frames of input feature expansion. The number of pre-training epochs for each RBM [41] layer was 1. The learning rate of pre-training was 0.4. As for fine-tuning, the learning rate and the maximum number of epochs were 0.00008 and 30, respectively. The mini-batch size was set to 128. The configuration parameters were chosen according to a previous investigation on speech enhancement [14]. Input and target features of DNN were globally normalized to zero mean and unit variance [15].

1) Evaluation With 40-Hour Training Data: The DNN models were first trained by 40 hours of training utterances simulating reverberant conditions. Fig. 4 displays the spectrograms of a test utterance, labeled "A," at RT60 = 0.6 s. Clearly HWW-DNN (see Fig. 4(c)), with blurred harmonics in the target normalized log-power spectrum, could not restore most of the anechoic spectrogram at intermediate frequencies, resulting in a little PESQ improvement of 0.22 relative to unprocessed reverberant speech. As for the proposed DNN with observable harmonics in the target normalized spectrum, the intermediate-frequency contents of the enhanced spectrogram (see Fig. 4(d)) were noted

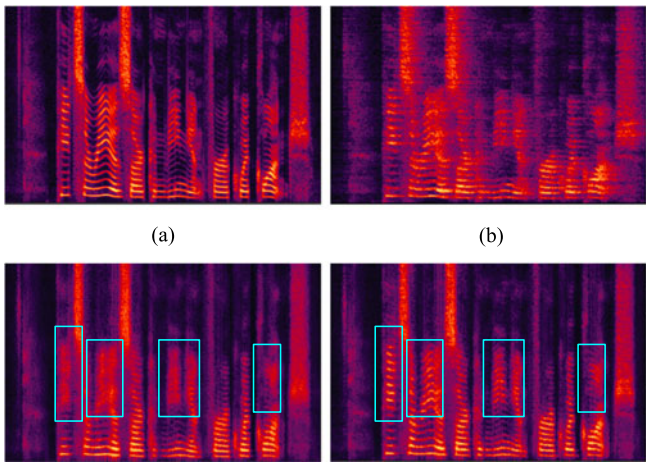


Fig. 4. Spectrograms of test utterance A, at RT60 = 0.6 s.: (a) anechoic speech (PESQ = 4.50), (b) reverberant speech (PESQ = 2.02), (c) HWW-DNN, (PESQ = 2.24), (d) proposed DNN (PESQ = 2.59).

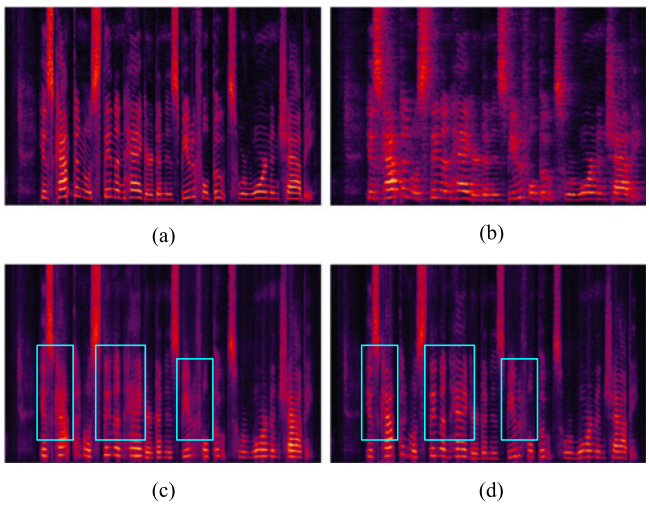


Fig. 5. Spectrograms of test utterance B, at RT60 = 0.6 s.: (a) anechoic speech (PESQ = 4.50), (b) reverberant speech (PESQ = 2.05), (c) HWW-DNN (PESQ = 1.80), (d) proposed DNN (PESQ = 2.47).

to be a closer match to the original anechoic speech spectrogram. And this enhanced signal achieved a considerable PESQ increase of 0.57.

Fig. 5 shows the spectrograms of another test utterance, labeled "B,', at RT60 = 0.6 s. Compared with reverberant speech, this time HWW-DNN (see Fig. 5(c)) even downgraded PESQ from 2.05 to 1.80, while the proposed DNN (see Fig. 5(d)) still boosted PESQ by 0.42 to 2.47. The results indicated that our DNN had a stable dereverberation performance.

With 40 hours of training data, the average PESQ, fwSegSNR and STOI results, of HWW-DNN and the proposed DNN on the test set at different RT60s, were illustrated in Fig. 6. When compared to unprocessed reverberant speech, the proposed DNN-based dereverberation system could achieve significant PESQ, fwSegSNR and STOI improvements of 0.49, 3.4 dB and

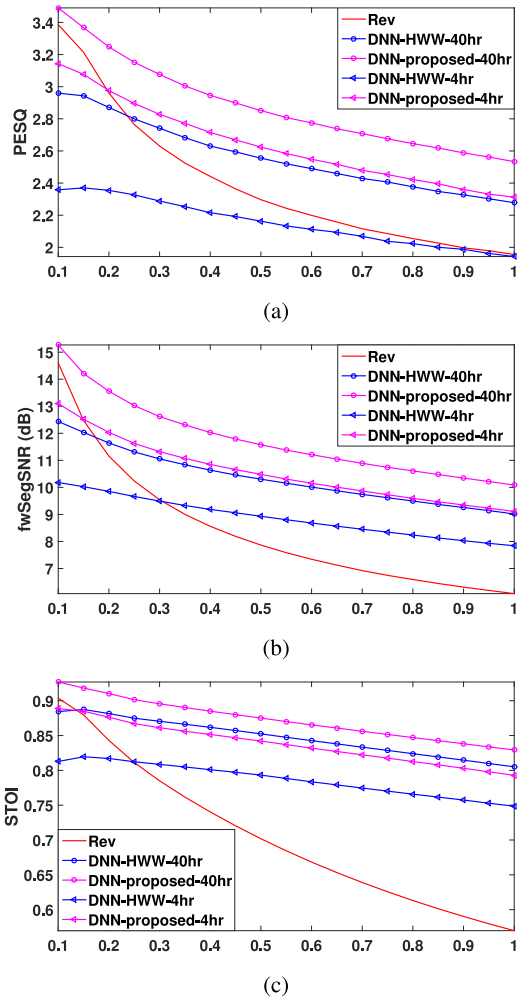


Fig. 6. Average PESQ, fwSegSNR and STOI of HWW-DNN and the proposed DNN with different training data size, on the test set at different RT60s.

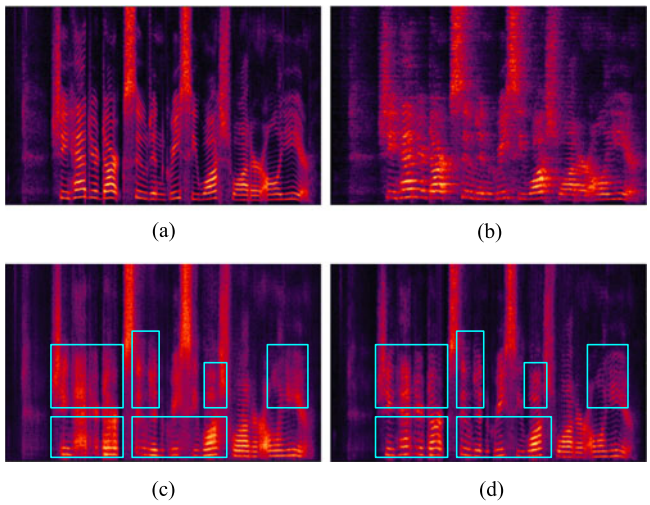


Fig. 7. Spectrograms of test utterance C, at RT60 = 0.6 s.: (a) anechoic speech (PESQ = 4.50), (b) reverberant speech (PESQ = 1.91), (c) HWW-DNN (PESQ = 1.42), (d) proposed DNN (PESQ = 2.21).

0.17, respectively, on the average at all RT60s, including mismatched conditions of RIRs and unseen speakers. This results demonstrated that the proposed DNN system had both powerful regression and generalization capabilities.

Furthermore, when compared with HWW-DNN, our DNN could significantly boost PESQ by 0.31 on the average at all RT60s, while achieving fwSegSNR and STOI improvements of 1.4 dB and 0.03, respectively. Another advantage of the proposed DNN was that it could substantially improve the speech quality in terms of the three kinds of objective measures at all RT60s. However, when compared with reverberant speech at low RT60s, the results of HWW-DNN started to show some considerable PESQ and fwSegSNR decreases and a little STOI decrement.

2) Evaluation With 4-Hour Training Data: We also trained the DNNs with about 4 hours of training data, only one tenth of the size of the training set used in Section IV-A1. Fig. 7 presents the spectrograms of another test utterance "C" again at RT60 = 0.6 s. With no sufficient training samples, the low and middle frequencies of HWW-DNN enhanced spectrogram (see Fig. 7(c)) were both blurred. Nevertheless, they were mostly restored by the proposed DNN system (see Fig. 7(d)), obtaining a good PESQ increase of 0.79 against HWW-DNN.

The average PESQ, fwSegSNR and STOI scores with 4 hours of training data were also given in Fig. 6. With a small training set size, HWW-DNN turned out to downgrade the PESQ scores at all RT60s. At RT60 = 0.1 s, the PESQ degradation was even more than 1. On the other hand the proposed DNN model still could improve the PESQ by 0.32 on the average at RT60 $\geq$ 0.25 s against reverberant speech, demonstrating the robustness of our DNN model even with no sufficient training samples. In terms of fwSegSNR and STOI, although the proposed model and HWW-DNN degraded speech for conditions below RT60 = 0.1 s and RT60 $\leq$ 0.25 s, respectively, the proposed system was superior to HWW-DNN at each RT60, achieving average fwSegSNR and STOI increases of 1.7 dB, and 0.05, respectively.

Note that, our DNN trained by 4 hours of training utterances outperformed HWW-DNN trained by 40 hours of training data at all RT60s, according to PESQ and fwSegSNR metrics.

B. Reverberation-Time-Aware DNN (RTA-DNN)

The following experimental settings were the same as in Section IV-A. The established large training set that consisted of about 40 hours of reverberant speech, was used to train DNN models.

In Sections IV-B1 and IV-B2, our proposed DNN models were trained to estimate the two parameters, R and N, needed to achieve top performances for each RT60, which was assumed to be known (oracle) in the dereverberation stage.

1) Frame-Shift-Aware DNN (FSA-DNN Oracle): To study the frame shift's effects on the dereverberation performance, a number of conventional frame-shift-independent DNNs, whose training and test utterances were enframed by different frame shift sizes with 7 frames of input feature expansion, are presented in Table I, denoted as "DNN-2ms," "DNN-4ms," etc. The numbers in bold denote maximum PESQ scores at

TABLE I
PESQ OF DIFFERENT FRAME SHIFTS AT DIFFERENT RT60S

RT60 (s)	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
Rev	3.39	2.96	2.63	2.44	2.30	2.20	2.12	2.05	2.00	1.96
DNN-2ms	**3.72**	3.33	3.09	2.93	2.80	2.70	2.59	2.50	2.42	2.34
DNN-4ms	3.71	**3.37**	3.14	2.98	2.86	2.75	2.66	2.57	2.49	2.42
DNN-8ms	3.68	3.36	**3.17**	**3.01**	**2.90**	**2.80**	**2.73**	**2.66**	2.58	2.51
DNN-16ms	3.49	3.25	3.08	2.95	2.85	2.77	2.71	2.65	**2.59**	**2.53**
DNN-24ms	3.08	2.90	2.77	2.67	2.60	2.54	2.49	2.44	2.40	2.35
DNN-32ms	2.40	2.41	2.36	2.29	2.25	2.21	2.17	2.14	2.10	2.06
FSA-DNN-oracle	3.75	3.39	3.16	3.03	2.91	2.82	2.74	2.66	2.59	2.52

TABLE II
PESQ OF VARIOUS FRAME EXPANSIONS AT DIFFERENT RT60S

RT60 (s)	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
Rev	3.39	2.96	2.63	2.44	2.30	2.20	2.12	2.05	2.00	1.96
DNN-1frame	3.37	3.04	2.81	2.63	2.48	2.36	2.27	2.17	2.08	1.99
DNN-3frame	3.50	3.23	3.03	2.88	2.76	2.66	2.58	2.50	2.43	2.35
DNN-5frame	**3.52**	**3.25**	3.06	2.93	2.83	2.74	2.66	2.60	2.54	2.48
DNN-7frame	3.49	3.25	**3.08**	**2.95**	2.85	2.77	2.71	2.65	2.59	2.53
DNN-9frame	3.46	3.24	3.06	2.94	**2.86**	2.77	2.71	2.65	2.59	2.54
DNN-11frame	3.44	3.22	3.05	2.94	2.86	**2.79**	**2.72**	**2.67**	**2.61**	**2.56**
ACA-DNN-oracle	3.52	3.25	3.06	2.94	2.84	2.78	2.72	2.67	2.62	2.56

each RT60 among all the conventional frame-shift-independent DNNs. "Rev" represents unprocessed reverberant speech.

Clearly the performances were affected greatly by the frame shift as these bold numbers show the optimal frame shift varies with RT60s. For conditions of $R \leq 16$ ms (half of frame length), "DNN-16ms" (conventional frame step) was superior to "DNN-2 ms," "DNN-4 ms" and "DNN-8 ms" for RT60 $\geq$ 0.4 s, 0.6 s and 0.9 s, respectively. In this case, lower frame shifts could not obtain better performances, even at the price of increased computational complexities. This results were consistent with our analysis in Section III-A. For conditions of $R > 16$ ms, "DNN-24 ms" and "DNN-32 ms" did not improve the PESQ scores for RT60 $\leq$ 0.2 s and 0.5 s, respectively, when compared with reverberant speech. It was not surprising because dereverberated speech could not be reconstructed exactly at an insufficient frame rate that caused aliasing. In addition, similar results of the frame shift's impact on the DNN-based dereverberation performance were obtained from other frame numbers of input feature expansion.

Inspired by the findings in Table I, a FSA-DNN was established by incorporating multi-resolution techniques. In the training stage, the utterances at distinct RT60s were enframed by an optimal frame shift that was extracted from these frame-shift-independent DNNs. In the dereverberation stage, as the RT60s were assumed to be known, the test data could be processed in the same manner as the training utterances.

As shown in Table I, relative to all the conventional frame-shift-independent DNNs, the proposed FSA-DNN ("FSA-DNN-oracle") yielded the highest PESQ scores at almost all RT60s, illustrating its powerful ability to adapt to the reverberant conditions. To be specific at RT60 $=$ 0.1 s, a PESQ improvement from 3.49 to 3.75 was obtained against "DNN-16 ms." While at RT60 $=$ 1 s, an increase from 2.34 to 2.52 was achieved against "DNN-2 ms."

2) Acoustic-Context-Aware DNN (ACA-DNN Oracle): To investigate context's impact on the quality of dereverberated speech, a collection of conventional acoustic-context-independent DNNs, which were fed with different numbers of frames expansion at a common frame shift of 16 ms (half of frame length), were shown in Table II, denoted as "DNN-1frame," "DNN-3frame," etc. Bold numbers denote the maximum PESQ scores at each RT60 among all the acoustic-context-independent DNNs. "Rev" represents reverberant speech.

TABLE III
FRAME SHIFT AND ACOUSTIC CONTEXT SIZES FOR DIFFERENT RT60S

RT60 (s)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
R (ms)	2	4	8	8	8	8	8	8	8	8
N	7	9	9	11	11	11	11	11	11	11

Table II shows that the context had an influence on PESQ because of the variation of the optimal context size with RT60s. PESQ for "DNN-5frame," "DNN-7frame," "DNN-9frame," "DNN-11frame" were in a strictly descending order at RT60 $=$ 0.1 s, even though the difference between them was small. It was surprising that more context information could not obtain higher quality enhanced speech. This could be explained by the theoretical analysis in Section III-B. While in a severely reverberant condition (e.g., RT60 $=$ 1 s), due to the strong frame-wise temporal correlation, this order turned to be ascending. Furthermore, a serious context information loss could make the enhanced speech quality in "DNN-1frame" and "DNN-3frame" worse than those of any other acoustic-context-independent DNNs at almost all RT60s.

With the findings in Table II, an ACA-DNN was designed to take advantage of the inter-frame correlation variation with the reverberant conditions. The optimal number of frame expansion at each RT60 was extracted from these acoustic-context-independent DNNs and then utilized to obtain normalized input features at different sizes. Due to the limitation that DNN could only take fixed length vectors as input, the normalized input features of unequal sizes were extended to 11 frames expansion by symmetrically padding zeros to the beginning and the end of the input vectors.

As shown in Table II, even though with a slight improvement compared with the acoustic-context-independent DNNs, the designed ACA-DNN ("ACA-DNN-oracle") still achieved the best PESQ scores at almost all RT60s, demonstrating its strong environmental robustness.

In the following sections we further explore two RTA-DNN systems that consider both effects by adopting the "optimal" RT60-dependent frame shift (R) and acoustic context (N) values in Table III, determined by experiments as Tables I and II but

TABLE IV
PESQ OF THREE ENVIRONMENT-AWARE DNNS AT DIFFERENT RT60S

RT60 (s)	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
FSA-DNN-oracle	3.75	3.39	3.16	3.03	2.91	2.82	2.74	2.66	2.59	2.52
ACA-DNN-oracle	3.52	3.25	3.06	2.94	2.84	2.78	2.72	2.67	2.62	2.56
RTA-DNN-oracle	3.77	3.41	3.17	3.05	2.95	2.87	2.80	2.73	2.66	2.61

with a full grid search[1] on R and N. In Section IV-B3 (oracle) these key parameter values were adopted in the established RTA-DNN based training system shown in the upper panel along with known RT60 in the dereverberation system shown in the bottom panel in Fig. 3 to obtain an upper-bound performances. Finally in Section IV-B4 (non-oracle), the RTA-DNN with an estimated RT60 for practical operational situations is introduced.

3) Reverberation-Time-Aware DNN (RTA-DNN Oracle): Assuming RT60 is known for each test utterance, a pair of R and N can be chosen from Table III to be utilized in the dereverberation stage in Fig. 3. The results in Table IV show that the "RTA-DNN-oracle" achieved the best performances at all known RT60s when compared with the top systems in Tables I and II, indicating that a combination of FSA-DNN and ACA-DNN could further improve the DNN robustness.

4) Reverberation-Time-Aware DNN (Estimated RT60): Finally we present experimental results to assess the performance of RTA-DNN in practical situations: (i) an RT60 estimator [42] was utilized in the dereverberation stage; (ii) another 100 randomly selected test utterances were utilized to evaluate the robustness of our proposed environment-aware techniques; (iii) the estimated utterance-level RT60 was rounded up to the nearest value of the multiples of 0.1; and (iv) the traditional reverberation-time-unaware DNN model ("DNN-proposed-40hr") was considered the DNN baseline, because it adopted the conventional frame rate (i.e., frame shift = half of frame length). (v) The results of HWW-DNN were also given in "DNN-HWW-40hr."

It can be seen from Fig. 8 that the designed "RTA-DNN-nonoracle-40hr" system achieved better PESQ, fwSegSNR and STOI than the DNN baseline at all RT60s, including extremely weak (RT60 = 0.1 s) and severe (RT60 = 1 s) reverberant cases. It indicated that the proposed RTA-DNN is robust enough to handle the slightly and highly reverberant situations which, however, were difficult in previous algorithms [2], [5]. To be specific at RT60 = 0.1 s, compared with the unprocessed reverberant speech, our proposed RTA-DNN achieved considerable PESQ and fwSegSNR improvements of 0.37 and 2.8 dB, respectively, while the results in DNN baseline showed only a little PESQ and fwSegSNR increases of 0.15 and 0.9 dB. It is also good to know that the results are only slightly worse than RTA-DNN with RT60s known labeled as "RTA-DNN-oracle-40hr" when RT60s are small.

[1]For conventional DNNs, R and N are chosen from a frame shift array of [2 ms, 4 ms, 8 ms, 16 ms] and a context acoustic array of [5 frames, 7 frames, 9 frames, 11 frames], respectively. That results in 4×4 DNNs. Frame shifts of 24 and 32 ms are excluded because an insufficient frame rate leads to aliasing, as shown in Table I. Acoustic contexts of 1 and 3 frames are excluded because a serious context information loss causes severe degradation, as illustrated in Table II. The "globally optimal" R and N at each RT60 are mostly determined by the best PESQ scores among all these RT60-independent DNNs.

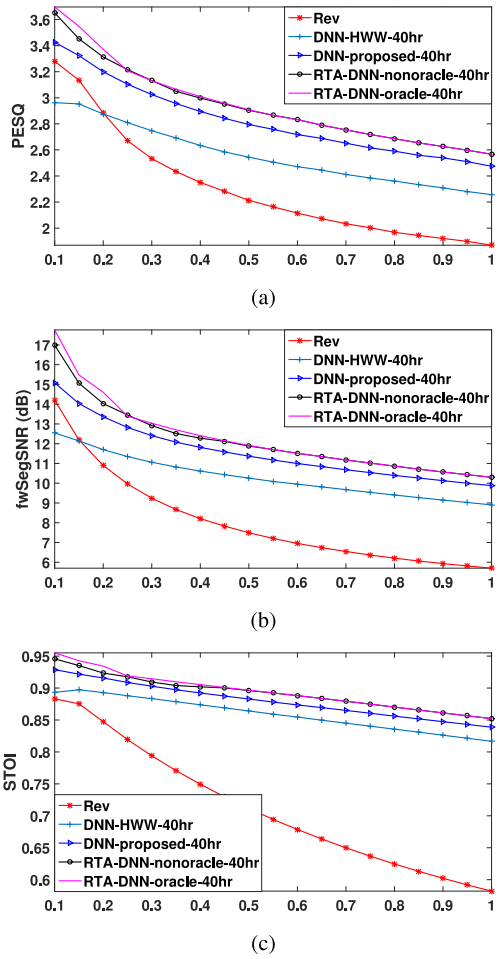


Fig. 8. Average PESQ, fwSegSNR and STOI results at different RT60s. "Rev," "RTA-DNN-nonoracle-40hr," "RTA-DNN-oracle-40hr," denote reverberant speech, RTA-DNN with estimated RT60 and RTA-DNN with RT60s known, with 40 hours of training data.

We notice that "DNN-proposed-4hr" in Fig. 6 degraded speech quality at low RT60s. As our proposed environment-aware framework showed good performance improvements, especially at short RT60s, we directly utilized the Table III obtained from 40 hours of training samples, to establish an RTA-DNN trained on 4 hours of reverberant data.

As shown in Fig. 9, compared with the DNN baseline, the designed "RTA-DNN-nonoracle-4hr" turned out to significantly improve the quality of reverberant speech at low RT60s. Specifically, the RTA-DNN achieved considerable PESQ, fwSegSNR and STOI increases of 0.43, 3.1 dB and 0.04, respectively, at RT60 = 0.1 s. Moreover, it outperformed the DNN baseline at all RT60s in terms of all three metrics, showing the robustness of the proposed environment-aware algorithm.

V. DISCUSSIONS ON GENERALIZATION CAPABILITIES

Since the DNN is a machine learning mechanism, it is important to evaluate its generalization capabilities to situations

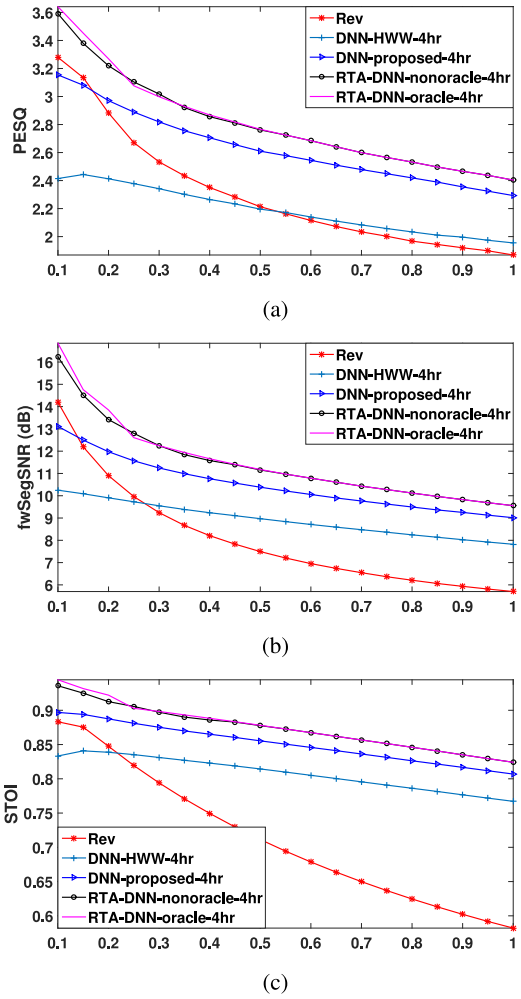


Fig. 9. Average PESQ, fwSegSNR and STOI results at different RT60s. "Rev," "RTA-DNN-nonoracle-4hr," "RTA-DNN-oracle-4hr," denote reverberant speech, RTA-DNN with estimated RT60 and RTA-DNN with RT60s known, with 4 hours of training data.

not seen in training. Therefore, we directly evaluate HWW-DNN system ("DNN-HWW"), our proposed DNN ("DNN-proposed") and RTA-DNN in non-oracle cases ("RTA-DNN") without retraining, in a series of mismatched conditions that are most commonly considered in practical applications. The DNN models were trained on 40 hours of training samples. PESQ, which has a high correlation with subjective evaluation scores [37], was used to evaluate the system performances.

A. Generalization to Room Sizes

The DNN systems, which were trained in the room of dimension 6 by 4 by 3 meters (length by width by height), were tested in a very different room of dimension 10 by 7 by 3 m, with the positions of loudspeaker and microphone unchanged. Three RT60s in the same room were generated by changing the absorption coefficients [8]. Fig. 10 shows the generalization results of the room size at different RT60s. Clearly, "DNN-proposed" yielded higher PESQ scores than the unprocessed reverberant speech and HWW-DNN at each RT60s. The

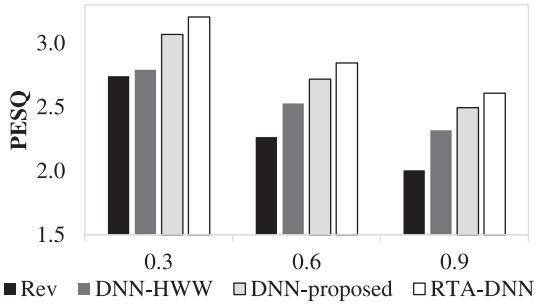


Fig. 10. Average PESQ results at different RT60s, tested in a new room.

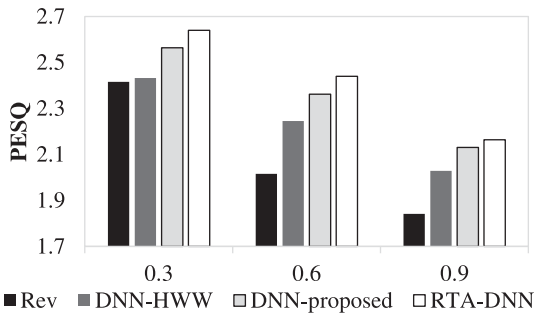


Fig. 11. Average PESQ results at different RT60s, tested at new loudspeaker and microphone positions.

results illustrate that although our proposed DNN model was only trained in a single room, it generalized well to an unseen room size. Moreover, "RTA-DNN" outperformed "DNN-proposed" at all RT60s, demonstrating the robustness of our proposed environment-aware approach to new room sizes. In addition, the "RTA-DNN" was also tested in other new room sizes and the results show that it constantly generalized well.

B. Generalization to Loudspeaker and Microphone Positions

In the training stage, the positions of the loudspeaker and the microphone were at (2, 3, 1.5) and (4, 1, 2) meters, respectively. In the dereverberation stage, we purposely changed the positions of the loudspeaker and the microphone to be at (1, 1.5, 2.5) and (3, 1, 0.5) meters, respectively. This time the room size was kept unchanged. Fig. 11 illustrates the generalization results of positions at different RT60s. Compared with unprocessed reverberant speech and HWW-DNN model, the proposed DNN significantly boosted PESQ scores at all RT60s. The "RTA-DNN" produced further improvements. It demonstrates that our proposed DNN and RTA-DNN had powerful generalization capabilities to loudspeaker and microphone positions. Moreover, the "RTA-DNN" was also tested in other new positions. We observe the same stable superior generalization capabilities in all tested conditions.

C. Generalization to Recorded RIRs

In order to show the DNN's generalization capabilities to real measured RIRs, without retraining, we directly evaluate the well-trained DNN systems in simulated rooms on the recorded

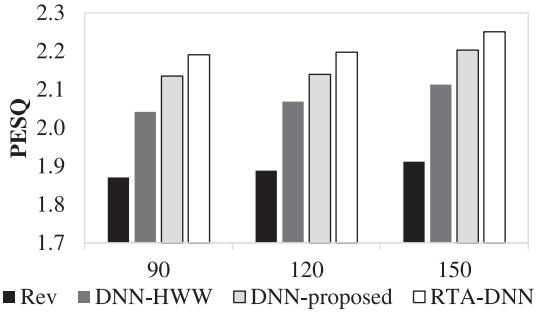


Fig. 12. Average PESQ results with different azimuth angles, tested on recorded RIRs.

RIRs [43]. The RIRs were measured in realistic rooms with various azimuth angles between head and source. As the data set offers binaural RIRs, we arbitrarily chose the left channel to generate reverberant speech. We chose the RIRs (RT60 = 1.94 s) in a stairway rather than regular rooms (e.g., meeting room, lecture room) to test the generalization capabilities of the DNN models, with azimuth angles of 90°, 120° and 150°. As shown in Fig. 12, the RTA-DNN model still could achieve the highest scores at each azimuth angle, illustrating that the proposed approach had a great generalization capability to measured RIRs in realistic rooms.

VI. Conclusion

In this paper, we have utilized a DNN-based regression model to enhance speech in reverberant conditions by adopting a linear output layer, with zero mean and unit variance normalization for target features. With a large or small training data, the proposed DNN feature mapping and normalization schemes always can improve the speech quality in terms of all objective performance metrics tested over the reverberant speech and HWW-DNN at each RT60, demonstrating the powerful regression capability and robustness of the proposed activation and normalization techniques.

We next show how the frame shift and acoustic context sizes affect the DNN-based dereverberation performance. An RTA-DNN was proposed by incorporating RT60-dependent frame shift and acoustic context parameters. The results show that the designed DNN model outperforms the conventional DNN approach at a wide range of RT60s, including extremely weak (RT60 = 0.1 s) and strong (RT60 = 1 s) reverberant situations, by taking into account the temporal resolution and frame-wise correlation at distinct RT60s. In addition, it generalizes well to unseen room size, loudspeaker and microphone positions, and recorded RIRs, which will yield significant benefits in many practical applications.

It should be noted that room reverberation and noise are two major causes for degradation of speech quality and intelligibility. We focus on the dereverberation problem in this paper, and would like to explore a DNN framework effective in both reverberant and noisy environments in future studies. Post-processing as in [5] and [44] will also be exploited. In addition, our proposed DNN dereverberation systems have been proven to be extremely effective in speech recognition tasks in [45].

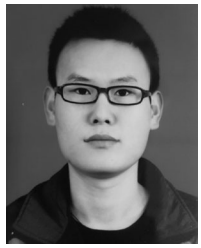
Acknowledgements

The authors would like to thank Dr. Y. Xu at University of Science and Technology of China, Dr. S. M. Siniscalchi at University of Enna Kore, and Dr. Z. Huang at Georgia Institute of Technology for their valuable suggestions.

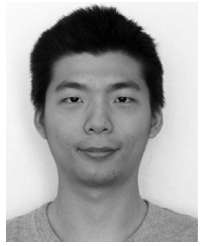
References

[1] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. London, U.K.: Springer, 2010.
[2] M. Wu and D. L. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 774–784, May 2006.
[3] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 534–545, May 2009.
[4] S. Mosayyebpour, M. Esmaeili, and T. A. Gulliver, "Single-microphone early and late reverberation suppression in noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 322–335, Feb. 2013.
[5] K. Han *et al.*, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.
[6] O. Schwartz, S. Gannot, and E. Habets, "An expectation-maximization algorithm for multi-microphone speech dereverberation and noise reduction with coherence matrix estimation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 9, pp. 1495–1510, Sep. 2016.
[7] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, no. 1, pp. 165–169, 1979.
[8] W. Sabine, *Collected Papers on Acoustics*. London, U.K.: Harvard Univ. Press, 1922.
[9] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.
[10] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech, Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.
[11] K. Kinoshita, T. Nakatani, and M. Miyoshi, "Fast estimation of a precise dereverberation filter based on speech harmonicity," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. 1073–1076.
[12] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[13] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[14] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.
[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.
[16] P. S. Huang *et al.*, "Deep learning for monaural speech separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 1562–1566.
[17] J. Du, Y.-H. Tu, Y. Xu, L.-R. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *Proc. Int. Conf. Signal Process.*, 2014, pp. 473–477.
[18] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4395–4399.
[19] K. Li, Z. Huang, Y. Xu, and C.-H. Lee, "DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech," in *Proc. Interspeech*, 2015, pp. 2578–2582.
[20] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4628–4632.
[21] B. Wu, K. Li, M. L. Yang, and C.-H. Lee, "A study on target feature activation and normalization and their impacts on the performance of DNN based speech dereverberation systems," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf., 2016*, to be published.
[22] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 4, pp. 679–681, Aug. 1982.

[23] 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.
[24] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. Workshop Autom. Speech Recog. Understanding*, 2011, pp. 1–4.
[25] P. Schniter, "Short-time Fourier transform," *Version*, vol. 2, no. 2005, p. 21, 1915.
[26] L. R. Rabiner and R. W. Schafer, *Theory and Application of Digital Signal Processing*. Upper Saddle River, NJ, USA: Prentice Hall, 2010.
[27] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Handbook of Speech Processing*. Berlin, Germany: Springer, 2008.
[28] C.-S. Jung, K. J. Han, H. Seo, S. S. Narayanan, and H.-G. Kang, "A variable frame length and rate algorithm based on the spectral kurtosis measure for speaker verification," in *Proc. Interspeech*, 2010, pp. 2754–2757.
[29] S. Mosayyebpour, H. Sheikhzadeh, T. A. Gulliver, and M. Esmaeili, "Single-microphone LP residual skewness-based inverse filtering of the room impulse response," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1617–1632, Jul. 2012.
[30] H. Kuttruff, *Room Acoustics*. Oxfordshire, U.K.: Spon Press, 2009.
[31] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Noida, India: Pearson, 2002.
[32] C. W. Gardiner, *Handbook of Stochastic Methods*. Berlin, Germany: Springer, 1985.
[33] S. Mosayyebpour, "Robust single-channel speech enhancement and speaker localization in adverse environments," Ph.D. dissertation, University of Victoria, Victoria, Canada, 2014.
[34] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.
[35] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 269–277, 2008.
[36] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," NIST, Tech. Rep., 1988.
[37] International Telecommunications Union's Telecommunication Standardization Sector, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Recommendation P.862, 2001.
[38] J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1978, pp. 586–590.
[39] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[40] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[41] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[42] A. Keshavarz, S. Mosayyebpour, M. Biguesh, T. A. Gulliver, and M. Esmaeili, "Speech-model based accurate blind reverberation time estimation using an LPC filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 6, pp. 1884–1893, Aug. 2012.
[43] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. 16th Int. Conf. Digital Signal Process.*, 2009, pp. 1–5.
[44] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in *Proc. Interspeech*, 2015, pp. 1508–1512.
[45] B. Wu, K. Li, Z. Huang, S. M. Siniscalchi, M. L. Yang, and C.-H. Lee, "A unified deep modeling approach to simultaneous speech dereverberation and recognition for the REVERB challenge," *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2017, submitted for publication.



Bo Wu received the B.Eng. degree in electronic information engineering from Southwest University, Chongqing, China, in 2012. He is currently working toward the Ph.D. degree in the National Laboratory of Radar Signal Processing, Xidian University, X'ian, China. From September 2014 to October 2016, he was a visiting student in the Center for Signal and Information Processing, Georgia Institute of Technology, Atlanta, GA, USA. His current research interests include signal processing, machine learning, and speech dereverberation.



Kehuang Li received the B.S. degree in information engineering and the M.S. degree in communication and information system from Shanghai Jiao Tong University, Shanghai, China, and the M.S. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, where he is currently working toward the Ph.D. degree in the Center of Signal and Information Processing, Department of Electrical and Computer Engineering. His research interests include signal processing, machine learning, and speech recognition.

Minglei Yang received the B.Eng. degree in electronic engineering and the Ph.D. degree in signal and information processing from Xidian University, Xi'an, China, in 2004 and 2009, respectively. Since June 2009, he has been with the National laboratory of Radar Signal Processing, Xidian University, where he is currently an Associate Professor. His current research interests include but not limited to signal processing, parameter estimation, and polarization information processing.

Chin-Hui Lee (F'97) is a professor at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Before joining academia in 2001, he had 20 years of industrial experience ending in Bell Laboratories, Murray Hill, NJ, USA, as a Distinguished Member of Technical Staff and the Director of the Dialogue Systems Research Department. He has published more than 400 papers and 30 patents, and was highly cited for his publications with an h-index of 64. He is a Fellow of the ISCA. 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 gave an ICASSP 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."