Learning Spectral Mapping for Speech Dereverberation and Denoising

Kun Han, *Student Member, IEEE*, Yuxuan Wang, *Student Member, IEEE*, DeLiang Wang, *Fellow, IEEE*, William S. Woods, *Member, IEEE*, Ivo Merks, *Member, IEEE*, and Tao Zhang, *Member, IEEE*

***Abstract*—In real-world environments, human speech is usually distorted by both reverberation and background noise, which have negative effects on speech intelligibility and speech quality. They also cause performance degradation in many speech technology applications, such as automatic speech recognition. Therefore, the dereverberation and denoising problems must be dealt with in daily listening environments. In this paper, we propose to perform speech dereverberation using supervised learning, and the supervised approach is then extended to address both dereverberation and denoising. Deep neural networks are trained to directly learn a spectral mapping from the magnitude spectrogram of corrupted speech to that of clean speech. The proposed approach substantially attenuates the distortion caused by reverberation, as well as background noise, and is conceptually simple. Systematic experiments show that the proposed approach leads to significant improvements of predicted speech intelligibility and quality, as well as automatic speech recognition in reverberant noisy conditions. Comparisons show that our approach substantially outperforms related methods.**

***Index Terms*—Deep neural networks (DNNs), denoising, dereverberation, spectral mapping, supervised learning.**

I. Introduction

In real-world environments, the sound reaching the ears comprises the original source (direct sound) and its reflections from various surfaces. These attenuated, time-delayed reflections of the original sound combine to form a reverberant signal. In reverberant environments, speech intelligibility is degraded substantially for hearing impaired listeners [19], and normal hearing listeners when reverberation is severe [32]. In addition, room reverberation when combined with background noise is particularly disruptive for speech perception. Reverberation and noise also cause significant performance degradation in automatic speech recognition (ASR) [17] and speaker identification systems [33], [43]. Given the prevalence of reverberation and noise, a solution to the dereverberation and denoising problems will benefit many speech technology applications.

Reverberation corresponds to a convolution of the direct sound and the room impulse response (RIR), which distorts the spectrum of speech in both time and frequency domains. Thus, dereverberation may be treated as inverse filtering. The magnitude relationship between an anechoic signal and its reverberant version is relatively consistent in different reverberant conditions, especially within the same room. Even when reverberant speech is mixed with background noise, it is still possible to restore speech to some degree from the mixture, because speech is highly structured. These properties motivate us to utilize supervised learning to model the reverberation and mixing process.

In this paper, we propose to learn the spectral mapping from reverberant speech to its anechoic version. The mapper is trained where the input is the spectral representation of reverberant speech and the desired output is that of anechoic speech. We then extend the spectral mapping approach to perform both dereverberation and denoising.

Deep neural networks (DNNs) have shown strong learning capacity [8]. A stacked denoising autoencoder (SDA) [37] is a deep learning method, and it can be trained to reconstruct the raw clean data from the noisy data, where hidden layer activations are used as learned features. Although SDAs were proposed to improve generalization, the main idea behind SDAs motivated us to utilize DNNs to learn the mapping from the corrupted data to clean data. A recent study [39] used DNNs to denoise acoustic features in each time-frequency unit for speech separation. In addition, Xu *et al.* [42] proposed a regression based DNN method for speech enhancement. Unlike these studies, our approach deals with reverberant and noisy speech, which is a substantially more challenging task. We note that an earlier version of our study dealing with just reverberation is published in [6] (more on this in Section V).

The paper is organized as follows. In the next section, we discuss related speech dereverberation and denoising studies. We then describe our approach in detail in Section III. The experimental results are shown in Section IV. We discuss related issues and conclude the paper in the last section.

Manuscript received August 25, 2014; revised December 14, 2014; accepted March 10, 2015. Date of publication March 25, 2015; date of current version April 10, 2015. This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-12-1-0130, a contract from Starkey, and the Ohio Supercomputer Center. A preliminary version of this work was presented at ICASSP 2014 [6]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rongshan Yu.

K. Han was with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA. He is now with Facebook, Menlo Park, CA 94025 USA (e-mail: hank@cse.ohio-state.edu).

Y. Wang is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: wangyuxu@cse.ohio-state.edu).

D. L. Wang is with the Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cse.ohio-state.edu).

W. S. Woods, I. Merks, and T. Zhang are with Starkey Hearing Technologies, Eden Prairie, MN 55344 USA (e-mail: bill_woods@starkey.com; ivo_merks@starkey.com; tao_zhang@starkey.com).

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.2015.2416653

II. Relation to Prior Work

Many previous approaches have been proposed to deal with speech dereverberation [24], [26]. Inverse filtering is one of the commonly used techniques [23]. Since the reverberation effect can be described as a convolution of clean speech with the room impulse response, the inverse filtering based approach first determines an inverse filter that can reverse the effects of the room response, and then estimates the anechoic signal by convolving the reverberant signal with the inverse filter. However, in many situations, the inverse filter cannot be determined directly and must be estimated, which is a hard problem. Further, this approach assumes that the RIR function is minimum-phase that is often not satisfied in practice [27]. Wu and Wang [41] utilized a two-stage approach including inverse filtering and spectral subtraction to deal with early reverberation and late reverberation separately, which relies on an accurate estimate of the inverse filter in one microphone scenarios. Other studies dealt with dereverberation by exploiting the properties of speech such as modulation spectrum [2], power spectrum [21], and harmonic structure [40], [30].

Recent studies show that the ideal binary mask (IBM) can be extended to suppress reverberation and improve speech intelligibility [19], [31], [32]. The IBM based approaches treat the direct sound or direct sound plus the early reflections as the target and the rest as the masker, and the dereverberated signal is resynthesized from the binary mask. Therefore, the IBM can still be considered as an effective computational goal for dereverberation. Hazrati *et al.* [7] proposed to estimate a binary mask based on a single variance-based feature against an adaptive threshold and yielded intelligibility improvements for cochlear implantees. In principle, the IBM based approach can deal with both reverberation and noise simultaneously; however, few previous studies aim to estimate the IBM for both dereverberation and denoising. Jin and Wang [14] use a multilayer perceptron to estimate the IBM for speech separation but the target is the reverberant noise-free speech.

III. Algorithm Description

We describe the algorithm in this section, including three subsections: feature extraction, model training, and post-processing.

A. Spectral Features

We first extract features for spectral mapping. Given a time domain input signal $s(t)$, we use the short time Fourier transform (STFT) to extract features. We first divide the input signal into 20-ms time frames with 10-ms frame shift, and then apply fast Fourier transform (FFT) to compute log spectral magnitudes in each time frame. For a 16 kHz signal, we use 320-point FFT and therefore the number of frequency bins is 161. We denote the log magnitude in the kth frequency and the mth frame as $X(m,k)$. Therefore, in the spectrogram domain, each frame can be represented as a vector $\mathbf{x}(m)$:

$$\mathbf{x}(m) = [X(m,1), X(m,2), \ldots, X(m,161)]^T \qquad (1)$$

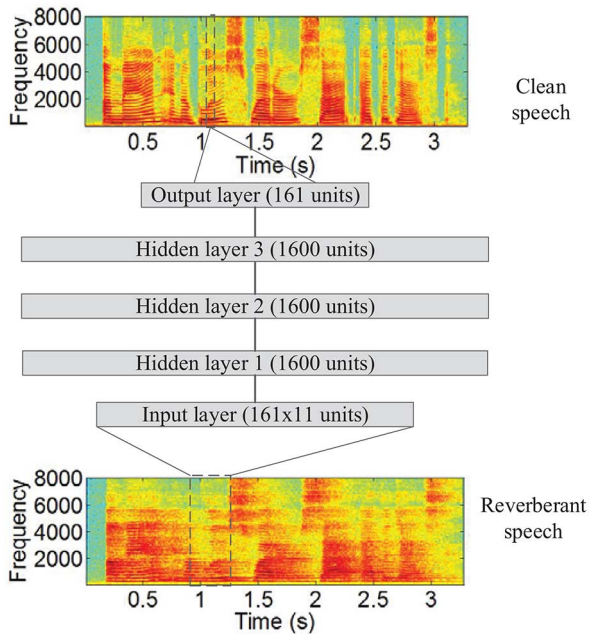

Fig. 1. (Color online) Structure of the DNN based spectral mapping. The inputs are the log spectra of the current frame and its neighboring frames, and the outputs are the log spectra of the current frame.

In order to incorporate temporal dynamics, we include the spectral features of neighboring frames into a feature vector. Therefore, the input feature vector for the DNN feature mapping is:

$$\tilde{\mathbf{x}}(m) = [\mathbf{x}(m-d), \ldots, \mathbf{x}(m), \ldots, \mathbf{x}(m+d)]^T \qquad (2)$$

where d denotes the number of neighboring frames on each side and is set to 5 in this study. So the dimensionality of the input is $161 \times 11 = 1771$.

The desired output of the neural network is the spectrogram of clean speech in the current frame m, denoted by a 161-dimensional feature vector $\mathbf{y}(m)$, whose elements correspond to the log magnitude in each frequency bin at the mth frame.

B. DNN Based Spectral Mapping

We train a deep neural network to learn the spectral mapping from reverberant, or reverberant and noisy, signals to clean signals.

The DNN in this study includes three hidden layers, as shown in Fig. 1. The input for each training sample is the log magnitude spectrogram in a window of frames $\tilde{\mathbf{x}}(m)$, and the number of input units is the same as the dimensionality of the feature vector. The output is the log magnitude spectrogram in the current frame $\mathbf{y}(m)$, corresponding to 161 output units. Each hidden layer includes 1600 hidden units. We use cross validation on a development set to train neural networks to choose the number of hidden layers and hidden units (see Section IV-B for more details).

The objective function for optimization is based on mean square error. Equation (3) is the cost for each training sample:

$$\mathcal{L}(\mathbf{y}, \mathbf{x}; \mathbf{\Theta}) = \sum_{c=1}^{C} (y_c - f_c(\mathbf{x}))^2 \qquad (3)$$

where $C = 161$ corresponds to the index of the highest frequency bin, $\mathbf{y} = (y_1, \ldots, y_C)^T$ is the desired output vector, and $f_c(\cdot)$ is the actual output of the cth neuron in the output layer. $\mathbf{\Theta}$ denotes the parameters we need to learn. To train the neural network, the input is normalized to zero mean and unity variance over all feature vectors in the training set, and the output is normalized into the range of [0,1]. The activation function in the hidden layers is the rectified linear function and the output layer uses the sigmoid function, shown in Eqs. (4) and (5) respectively:

$$f(x) = \max(0, x) \tag{4}$$

$$f(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

The weights of the DNN are randomly initialized without pretraining. We use backpropagation with a mini-batch of 512 samples for stochastic gradient descent to train the DNN model, and the cost function in each mini-batch is computed as the summation over multiple training samples using Eq. (3). The optimization technique uses gradient descent along with adaptive learning rates and a momentum term [3].

The output of DNN is the estimated log magnitude spectrogram of clean speech. With the capacity of learning internal representations, DNN promises to be able to encode the spectral transformation from corrupted speech to clean speech and help to restore the magnitude spectrogram of clean speech.

C. Post-Processing

After the DNN generates magnitude spectrogram estimates, we need to resynthesize time-domain signals using the inverse FFT process.

A straightforward method to reconstruct time-domain signals is to directly apply inverse the short-time Fourier transform (iSTFT) using the DNN-generated magnitude and the phase from unprocessed time-domain signals. However, the original phase of noise-free speech is corrupted, and the corruption usually introduces perceptual disturbances and leads to negative effects on sound quality. In addition, the STFT is computed by concatenating Fourier transforms of overlapping frames of a signal, and thus is a redundant representation of the time-domain signal. For a spectrogram-like matrix in the time-frequency domain, it is not guaranteed there exists a time-domain signal whose STFT is equal to that matrix [4], [20]. In other words, the magnitude spectrograms of the resynthesized time-domain signal could be different from the one we intended to resynthesize a signal from. This inconsistency should be taken into account for synthetic or modified spectrograms, like our DNN-generated magnitudes.

In order to minimize the incoherence between the phase and the magnitude from which we want to reconstruct a signal, we use an iterative procedure to reconstruct a time-domain signal as given in Algorithm 1 [4]:

Here, $N = 20$ in our study. The algorithm iteratively updates the phase ϕ at each step by replacing it with the phase of the STFT of its inverse STFT, while the target magnitude Y^0 is the DNN-generated output, which is always fixed. The iteration aims to find the closest realizable magnitude spectrogram consistent with the given magnitude spectrogram.

Algorithm 1 Iterative signal reconstruction

Input: Target magnitude Y^0, noisy phase ϕ^0 and iteration number N

Output: Time-domain signal s

$Y \leftarrow Y^0$, $\phi \leftarrow \phi^0$, $n \leftarrow 1$
while $n \leq N$ **do**
  $s^n \leftarrow \mathrm{iSTFT}(Y, \phi)$
  $(Y^n, \phi^n) \leftarrow \mathrm{STFT}(s^n)$
  $Y \leftarrow Y^0$
  $\phi \leftarrow \phi^n$
  $n \leftarrow n + 1$
end while
$s \leftarrow s^N$

We use the above post-processing to reconstruct a time-domain signal as a waveform output of our system.

Fig. 2 shows an example of the spectral mapping for a female sentence "A man in a blue sweater sat at the desk". Figs. 2(a) and (b) show the log magnitude spectrogram of the clean speech and the reverberant speech with $T_{60} = 0.6$ s. The corresponding DNN output is shown in Fig. 2(c). As shown in Fig. 2(c), the smearing energy caused by reverberation is largely removed or attenuated, and the boundaries between voiced and unvoiced frames are considerably restored, showing that the DNN output is a very good estimate of the spectrogram of the clean speech. Fig. 2(d) is the magnitude spectrogram of the time-domain signal resynthesized from the magnitude in Fig. 2(c) and reverberant phase. Comparing Figs. 2(c) with (d), the spectrogram in Fig. 2(d) is not as clean as the DNN output in Fig. 2(c) because of the use of reverberant phase and inconsistency of STFT. Fig. 2(e) is the spectrogram of the time-domain signal using post-processing, where the spectrogram is improved by iterative signal reconstruction.

IV. Experiments

A. Metrics and Parameters

We quantitatively evaluate our approach by two objective measurements of speech intelligibility: frequency-weighted segmental speech-to-noise ratio (SNR_{fw}) [22] and short-time objective intelligibility measure (STOI) [34]. Specifically, SNR_{fw} is a speech intelligibility indicator, computing a signal-to-noise estimate for each critical band:

$$\mathrm{SNR}_{fw} = 10M \sum_{m=1}^{M} \frac{\sum_{k=1}^{K} W(k) \log_{10} \frac{|S(m,k)|^2}{|S(m,k) - \hat{S}(m,k)|}}{\sum_{k=1}^{K} W(k)} \tag{6}$$

where $W(k)$ is the weight placed on the kth frequency band which is taken from the articulation index [13], K is the number of bands, M is the total number of frames in the signal, $S(m, k)$ is the critical-band magnitude of the clean signal in the kth frequency band at the mth frame, and $\hat{S}(m, k)$ is the corresponding spectral magnitude of the processed signal in the same band.

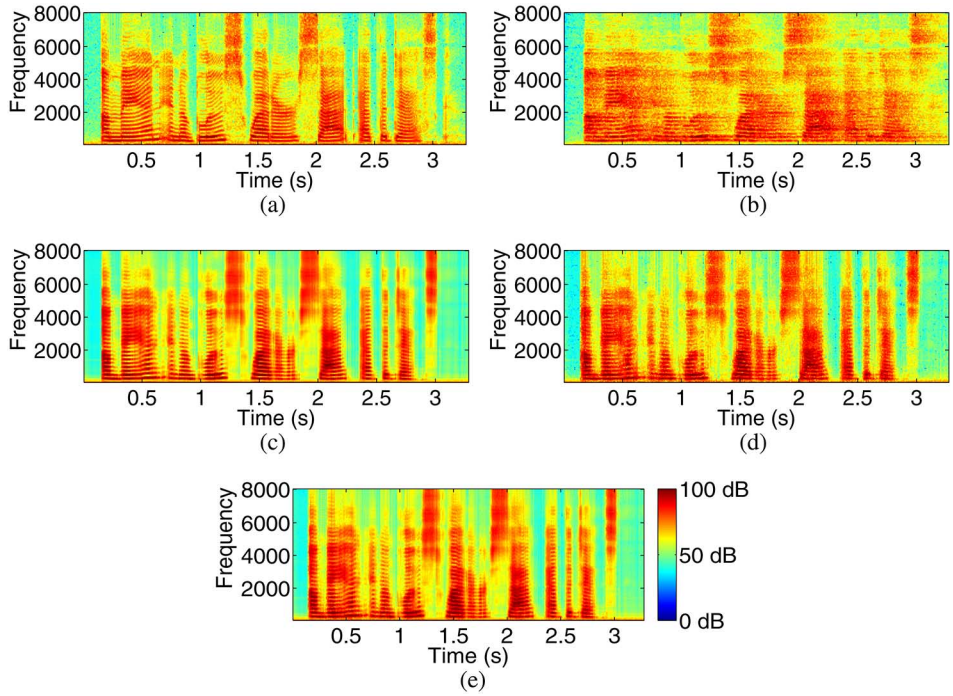

Fig. 2. (Color online) DNN dereverberation results. (a) Log magnitude spectrogram of clean speech. (b) Log magnitude spectrogram of reverberant speech with $T_{60} = 0.6$s. (c) DNN outputs. (d) Log magnitude spectrogram of resynthesized signal. (e) Log magnitude spectrogram of resynthesized signal with post-processing.

The STOI is recently proposed as an index of speech intelligibility. It computes the correlation between temporal envelopes of the clean and processed speech in short-time segments as an intelligibility indicator, ranging from 0 to 1. STOI has been shown to have high correlation with speech intelligibility of human listeners [34].

In addition, we evaluate speech quality using Perceptual Evaluation of Speech Quality (PESQ) [29], which computes disturbance between clean speech and processed speech using cognitive modeling as a speech quality score. The range of PESQ score is from -0.5 to 4.5.

As we mentioned in Section III-A, we utilize context information using a concatenation of features from 5 frames on each side of the current frame. Temporal information is an important property for speech signals, and thus adding these neighboring frames should be helpful to learn a spectral mapping. We have conducted experiments using different window sizes. Comparing with the 11-frame window, the SNR_{fw} results of a 7-frame window and a 3-frame window are worse by around 0.5 dB and 2.5 dB, respectively.

The architecture of the DNN influences its learning performance. We have conducted experiments using different numbers of hidden layers. A DNN with three hidden layers performs slightly better than that with two hidden layers in terms of SNR_{fw} (by 0.2 dB), and better than that with a single hidden layer (by 1.1 dB).

B. Dereverberation

We first evaluate dereverberation performance in this section. To mimic room acoustics, we generate a simulated room with a size of 10 m $\times$ 7 m $\times$ 3 m (length, width, height) and vary reflection coefficients to yield a specific T_{60} [5]. For each T_{60} condition, we simulate a set of RIRs by uniformly and randomly positioning a sound source and a receiver in the room, the distance between which is controlled to be greater than 0.5 m to avoid close-talking scenarios. To train the system, we use three reverberation times of 0.3, 0.6, and 0.9 s, and for each T_{60} we generate 2 different RIRs. We use 200 anechoic female utterances from the IEEE corpus [12] to form the training set. Therefore, there are $200 \times 3 \times 2 = 1200$ reverberant sentences in the training set. The test set includes 60 reverberant sentences, corresponding to 20 speech utterances, three T_{60}s, and one RIR. We also construct a development set including 10 utterances mixed with three RIRs with the three T_{60}s for cross validation in order to choose the parameters of the neural network. All the utterances in the IEEE corpus are from one female speaker. There is no overlap in the utterances and RIRs used in the training, development, and test sets.

We compare the proposed approach with two dereverberation algorithms. Hazrati *et al.*[7] recently proposed a dereverberation approach, utilizing a variance-based feature from the reverberant signal and comparing its value against an adaptive threshold to compute a binary mask for dereverberation. Wu and Wang [41] used estimated inverse filters and spectral subtraction to attenuate early reverberation and late reverberation, respectively. We also perform dereverberation using the IBM with the relative criterion -5 dB suggested in [32]. Since the IBM is generated from the anechoic speech against the reverberant speech, the results can be considered as a ceiling performance of binary masking systems.

In Fig. 3, we show the evaluation results in terms of frequency-weighted SNR, STOI, and PESQ, as well as those of the comparison systems. For SNR_{fw} results shown in Fig. 3(a), without iterative reconstruction post-processing, the DNN

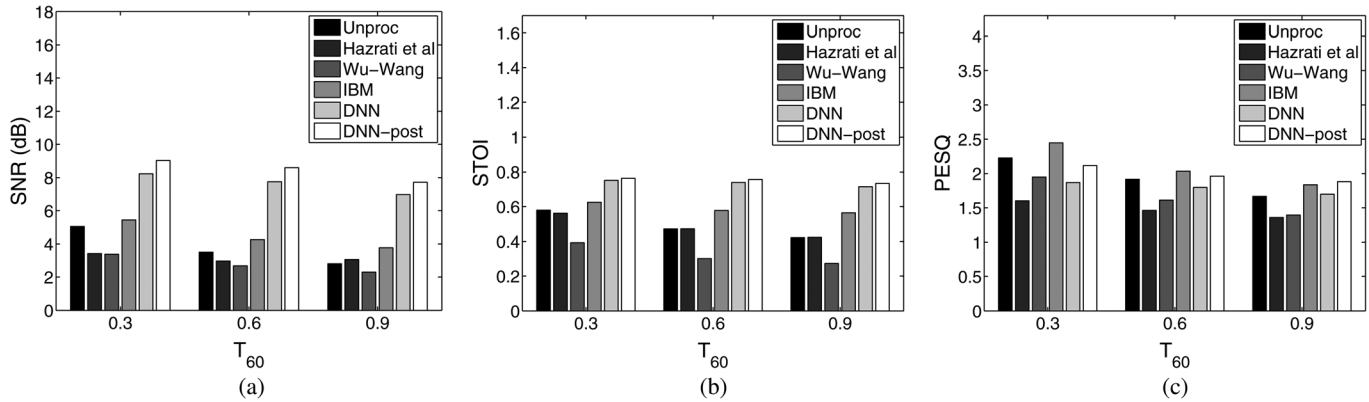

Fig. 3. DNN based dereverberation results: (a) SNR_{fw}, (b) STOI, (c) PESQ. "Unproc" denotes the results for unprocessed reverberant speech. 'Hazrati *et al.*" and "Wu-Wang" denote two baselines as described. "IBM" denotes the dereverberation results using the IBM. "DNN" denotes the proposed spectral mapping approach using reverberant phase without post-processing. "DNN-post" denotes the proposed spectral mapping approach with iterative signal reconstruction.

based approach improves SNR_{fw} relative to the unprocessed reverberant speech by 4 dB on average. The post-processing further boosts SNR_{fw} by around 1 dB. Comparing with Hazrati *et al.* and Wu and Wang, our DNN based methods achieve the highest SNR_{fw} scores. In addition, the proposed approach even generates better results than the IBM based approach. Consistent with SNR_{fw}, Fig. 3(b) shows that the proposed methods yield high STOI scores under each reverberation time, higher than the unprocessed and the other two approaches by more than 0.25. As shown in Fig. 3(c), the proposed approach does not boost PESQ scores for the conditions of $T_{60} \leq 0.6$ s, partly because mild reverberation does not lead to significant sound quality degradation. When the reverberation time is long, the PESQ score is boosted by our approach as shown in the condition of $T_{60} = 0.9$ s.

Since our approach is a supervised learning method, it is important to evaluate its generalizability. We generate another set of RIRs with T_{60} from 0.2 to 1.0 s, with the increment of 0.1 s. Note that, none of RIRs in this experiment are seen in the training set as they are created from different rooms. We compare unprocessed signals with our DNN based approach without post-processing. Fig. 4 shows the generalization results of SNR_{fw} for different T_{60}s. Compared with the unprocessed reverberant speech, the proposed approach substantially improves SNR_{fw} in each T_{60} and the advantage becomes increasingly larger as T_{60} increases, demonstrating that our approach generalizes well to new reverberant environments in a wide range. Fig. 4 also shows the DNN processed results for anechoic speech, corresponding to $T_{60} = 0$ s in the figure.

To further evaluate generalization of our approach, we conduct a cross-corpora experiment, i.e., we directly evaluate the system on the TIMIT corpus [44] without retraining. We randomly choose ten utterances from ten different female speakers (one utterance from one speaker) from the TIMIT corpus and generate reverberant signals using the same three RIRs used as in the above experiments. Fig. 5 shows the comparison results. As shown in the figure, although the DNN model is trained on the IEEE corpus with only one speaker, it generalizes well to another corpus with multiple speakers. The relative improvements are smaller, but our supervised learning based approach with

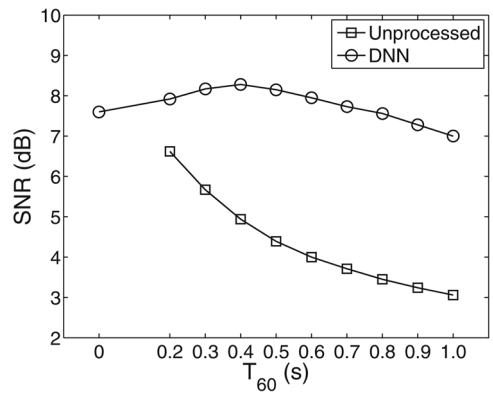

Fig. 4. Generalization results in different T_{60} s. "DNN" denotes the DNN based spectral mapping approach without post-processing, and "Unprocessed" the results for original reverberant speech.

no retraining outperforms other approaches in terms of PESQ, SNR_{fw}, and STOI.

Although mild to moderate reverberation does not significantly impact speech perception for normal hearing listeners, an adverse effect occurs when reverberation is severe [32]. We have also conducted dereverberation experiments for strong reverberation conditions, when T_{60} is greater than 1.0s. Similar to the above experiment, we use the same utterances to generate reverberant sentences with T_{60} set to 1.2 s, 1.5 s, and 1.8 s. The training and test sets use different utterances and different RIRs. Experimental results are shown in Fig. 6. Comparing with unprocessed sentences, the DNN based methods improve SNR_{fw} and STOI scores. Note that, unlike moderate reverberation conditions as shown in Fig. 3, PESQ scores are boosted in each reverberation time as shown in Fig. 6(c). In these conditions, the post-processing achieves consistently better performance for each metric.

The above experiments use simulated RIRs to generate reverberant signals. We now evaluate using recorded RIRs [11]. The RIRs were captured in real rooms with sound sources placed on the frontal azimuthal plane at different angles. Since the recordings consist of binaural RIRs, we choose (arbitrarily) the RIRs from the left ear to generate reverberant signals. We choose the RIRs in a medium-large sized seminar and presentation room

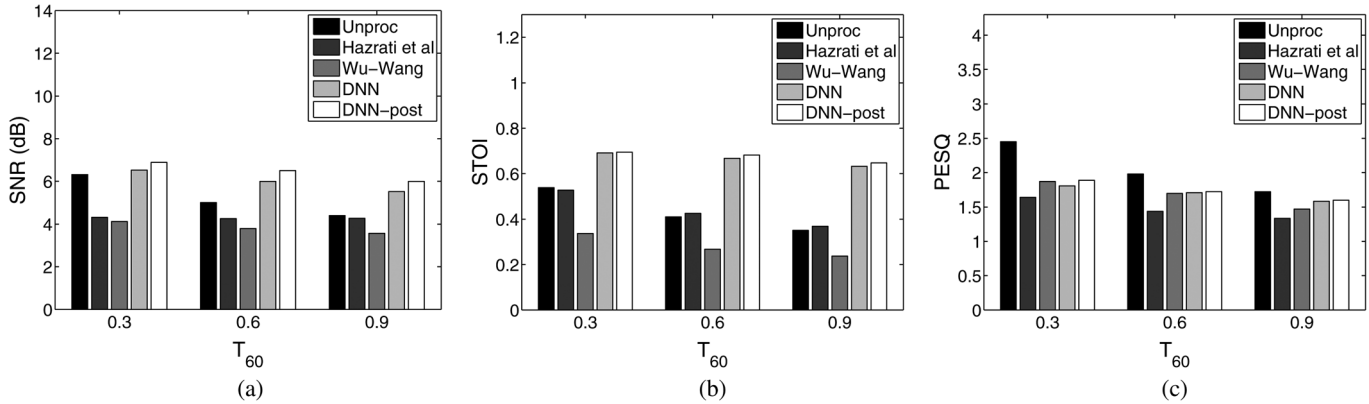

Fig. 5. Cross-corpora dereverberation results. The DNN model is trained on the IEEE corpus, but tested on the TIMIT corpus: (a) SNR_{fw}, (b) STOI, (c) PESQ. See Fig. 3 caption for notations.

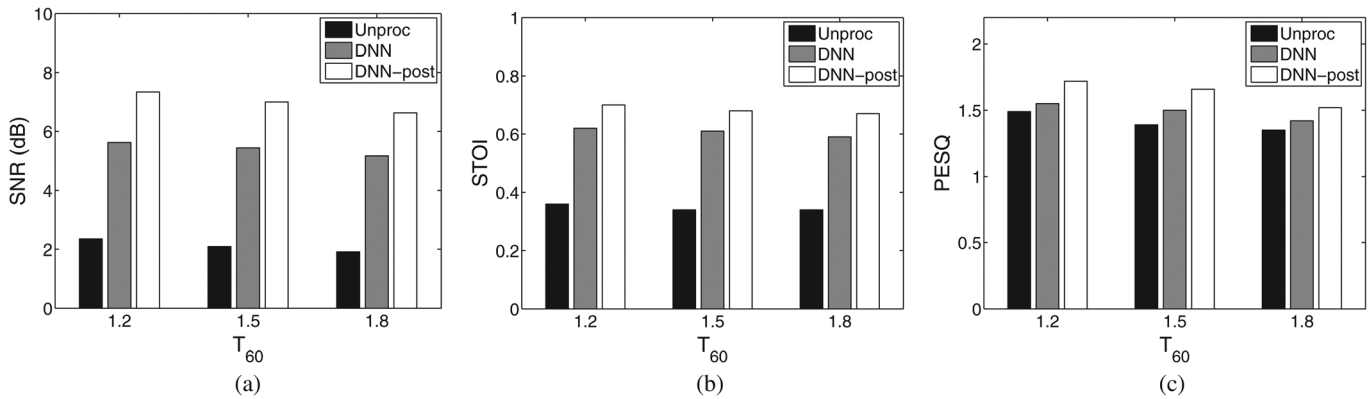

Fig. 6. DNN based dereverberation results under strong reverberation conditions: (a) SNR_{fw}, (b) STOI, (c) PESQ. See Fig. 3 caption for notations.

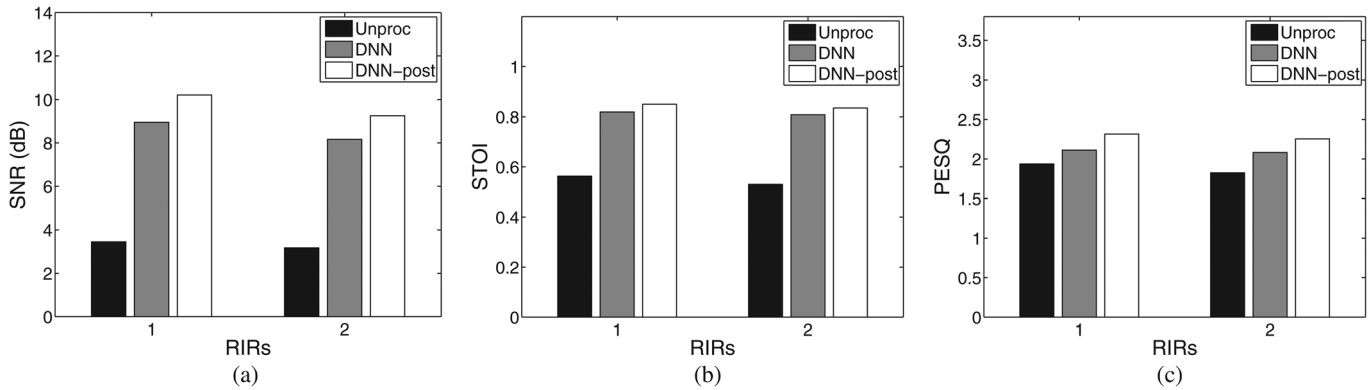

Fig. 7. DNN based dereverberation results using recorded RIRs: (a) SNR_{fw}, (b) STOI, (c) PESQ. RIR 1 corresponds to the azimuth angle of $-10°$, and RIR 2 the azimuth angle of $30°$.

with $T_{60} = 0.89$ s. We train the system using five RIRs with azimuth angles of $-40°$, $-20°$, $0°$, $20°$, and $40°$ and test using two RIRs with azimuth angles of $-10°$ and $30°$. As shown in Fig. 7, the DNN based approach boosts SNR_{fw} by more than 5 dB, STOI by around 0.3, and PESQ by 0.2. The post-processing step produces further improvements.

C. Dereverberation and Denoising

Our approach can deal with not only reverberation but also background noise. We can use the same supervised approach to perform dereverberation and denoising simultaneously. In this situation, the input to the neural network is the log magnitude spectrogram of reverberant and noisy speech, and the output is the log magnitude spectrogram of anechoic clean speech.

We conduct experiments for dereverberation and denoising. We generate a simulated room corresponding to a specific T_{60} and randomly create a set, $\{r_T, r_I, r_M\}$, representing the locations of the target, the interference and the microphone inside the room, respectively [14]. From these locations, a reverberant mixture $r(t)$ is constructed by

$$r(t) = h_T(t) * s(t) + \alpha h_I(t) * n(t) \tag{7}$$

where, $h_T(t)$ and $h_I(t)$ are the RIR of the target and the interference at the microphone location, respectively. "$*$" denotes

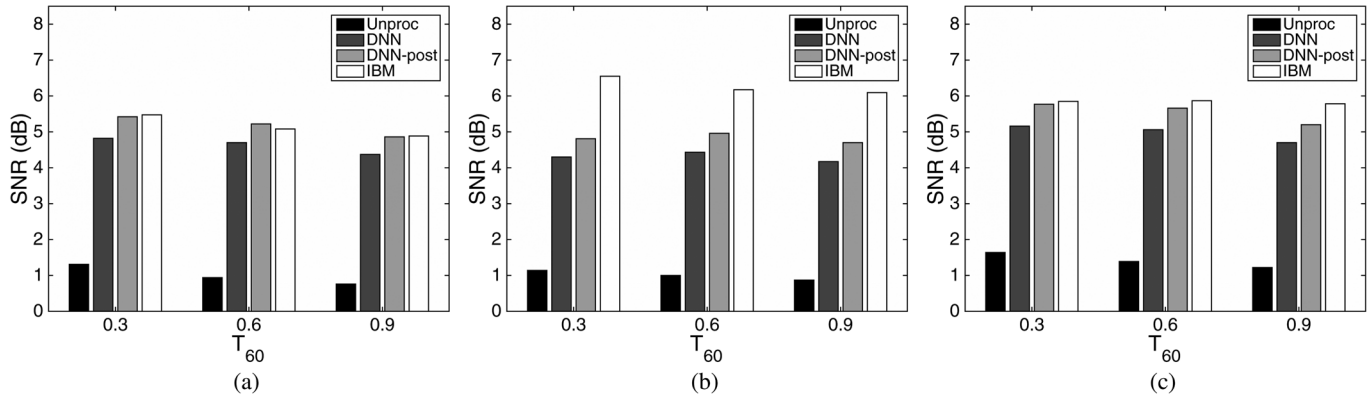

Fig. 8. SNR_{fw}s for seen noises: (a) babble noise, (b) factory noise, (c) speech-shaped noise. "Unproc" denotes the results for unprocessed reverberant speech. "DNN" denotes the proposed spectral mapping approach without post-processing. "DNN-post" denotes the proposed spectral mapping approach with iterative signal reconstruction processing. "IBM" denotes the results using the IBM.

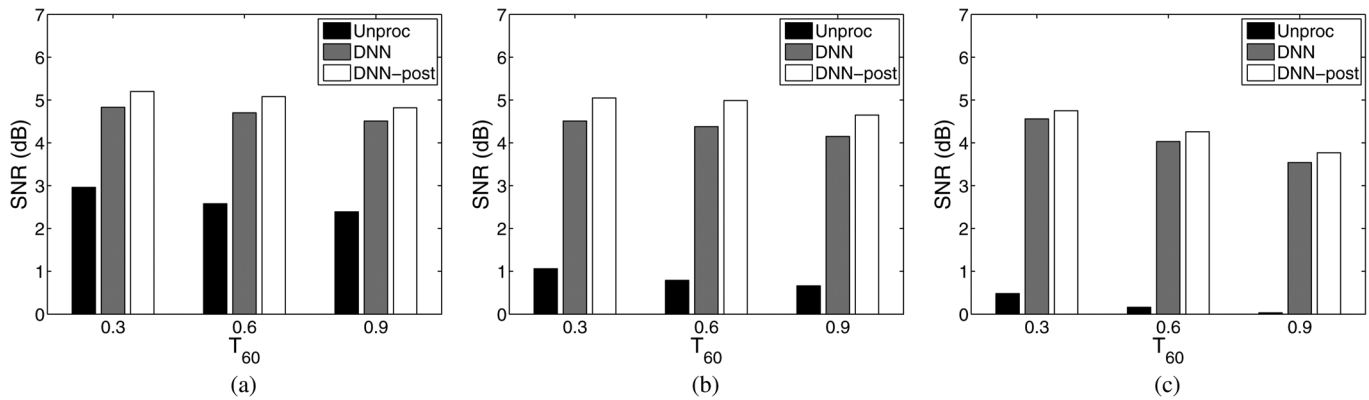

Fig. 9. SNR_{fw}s for new noises:: (a) white noise, (b) cocktail-party noise, (c) crowd noise in playground.

convolution. We use α as a coefficient to control the SNR of the mixture.

We simulate three acoustic rooms and their T_{60}s are 0.3, 0.6, and 0.9 s, respectively. The training set contains reverberant mixtures including 200 utterances convolved with 3 RIRs and mixed with 3 noise types: speech-shaped noise, factory noise and babble noise [16] at 0 dB SNR. Here, the SNR is computed as the ratio of the energy of the reverberant noise-free signal to that of the reverberant noise-only signal. To test the system, 60 new reverberant utterances are mixed with the three training noises and three new noises, white noise, cocktail party noise, and crowd noise in playground [10], under each T_{60} but using different RIRs.

Fig. 8 and Fig. 9 show SNR_{fw} results for seen noises and new noises, respectively. The DNN based method increases SNR_{fw} by 4.5 dB for seen noises and post-processing further yields 0.5 dB improvement. The proposed methods also achieve significant improvement for new noises, and the average advantage is around 3 dB, showing good generalization of the proposed approach. As a benchmark, we also apply the IBM for dereveberationa and denoising for seen noises, where the target in the IBM is the anechoic clean speech and the masker is the remaining signal. As shown in Fig. 8, the DNN based results are lower than the IBM results by around 0.5 dB on average, as the IBM utilizes ideal information.

STOI scores are shown in Fig. 10 and Fig. 11, and DNN and DNN with post-processing have similar performances. On average, both increase STOI scores by around 0.15 for seen noises and 0.13 for new noises. As expected, the IBM scores are the highest.

As shown in Fig. 12 and Fig. 13, PESQ results are improved by the proposed approach for both seen noises and new noises. For seen noises, the average PESQ scores for unprocessed, DNN, and DNN with post-processing sentences are 1.06, 1.31, 1.45, respectively. IBM processing leads to lower PESQ scores. For unseen noises, they are 1.13, 1.14, 1.21, respectively. These results demonstrate that the proposed approach improves speech quality when speech is corrupted by both noise and reverberation.

D. Robust Speech Recognition

The above evaluations show that our DNN based spectral mapping significantly attenuates reverberation and noise and produces good estimates of magnitude spectrogram of clean speech. As ASR algorithms only utilize magnitude spectrogram, one would expect our approach to improve ASR performance in reverberant and noisy conditions.

In this evaluation, we use the second CHiME challenge corpus (track 2) to evaluate ASR performance [36]. In the CHiME-2 corpus, the utterances are taken from the speaker-independent 5k vocabulary subset of the Wall Street Journal (WSJ0) corpus. Each utterance is convolved with one recorded binaural room impulse response corresponding to a front position at a distance of 2 m, and then mixed with binaural

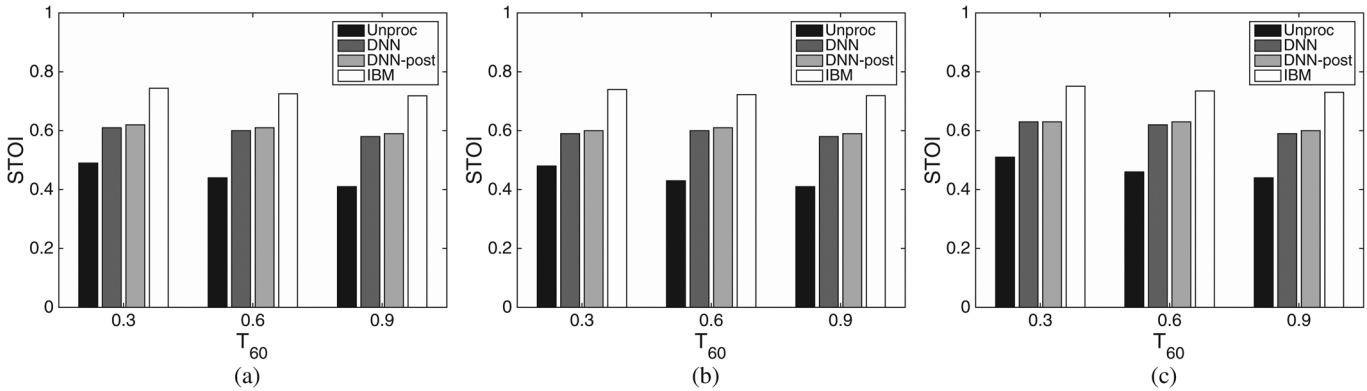

Fig. 10. STOI scores for seen noises: (a) babble noise, (b) factory noise, (c) speech-shaped noise.

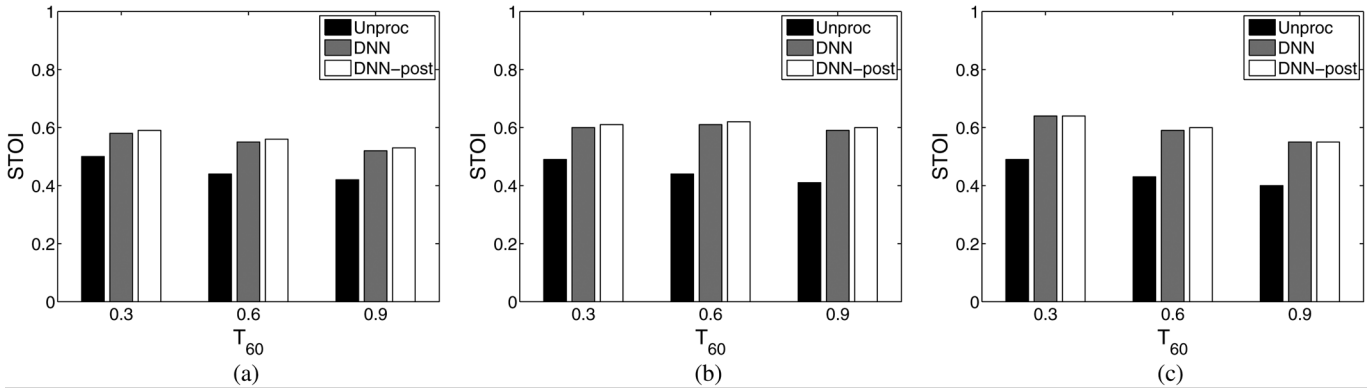

Fig. 11. STOI scores for new noises:: (a) white noise, (b) cocktail-party noise, (c) crowd noise in playground.

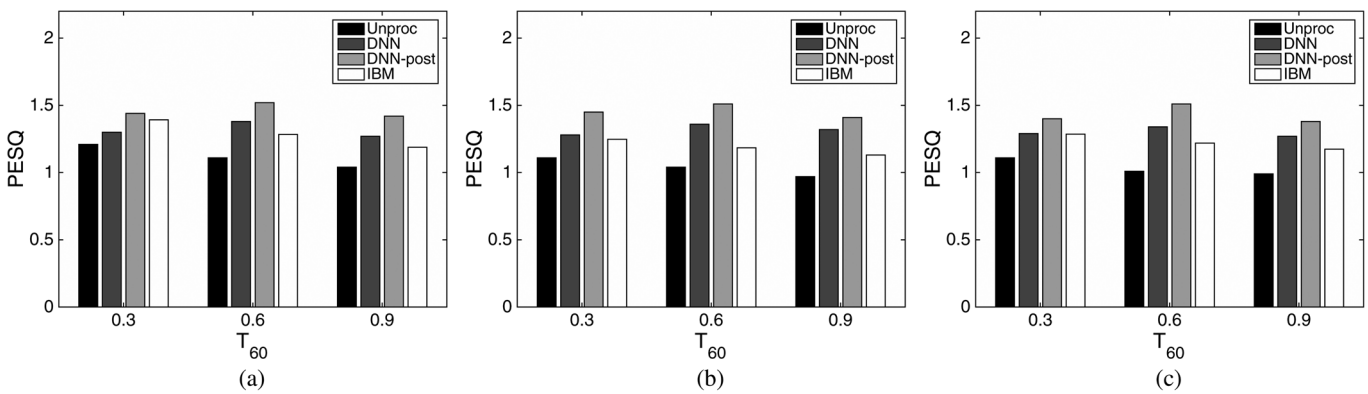

Fig. 12. PESQ scores for seen noises: (a) babble noise, (b) factory noise, (c) speech-shaped noise.

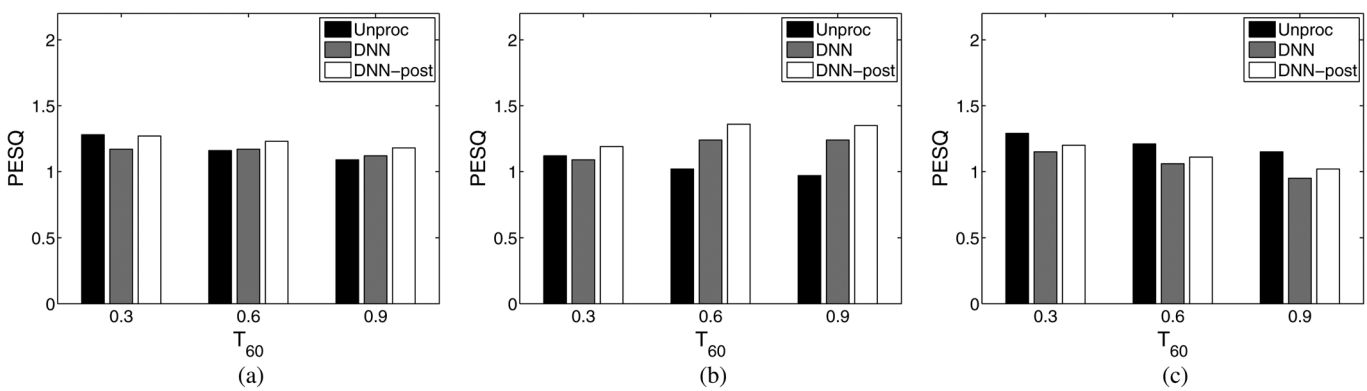

Fig. 13. PESQ scores for new noises:: (a) white noise, (b) cocktail-party noise, (c) crowd noise in playground.

recordings of real room noise over a period of days in the same family living room at 6 SNRs of -6, -3, 0, 3, 6, 9 dB. Since our study focuses on monaural speech processing, only single channel signals (left ear) are used.

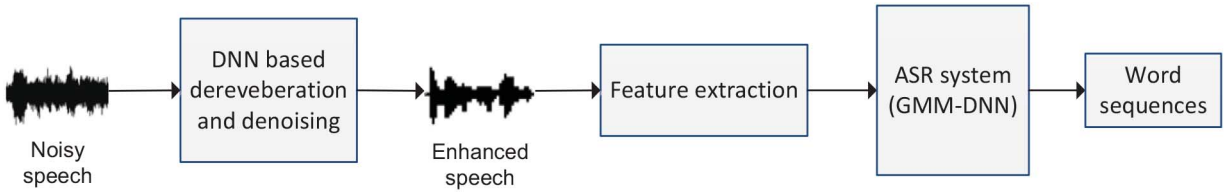

Fig. 14. Diagram of an ASR system with a DNN based front-end for dereverberation and denoising.

To perform ASR, the proposed approach is treated as a front-end to enhance all sentences in both training and test datasets as shown in Fig. 14. We first randomly choose 3000 out of 7138 sentences from the CHiME-2 training set to train our DNN based dereverberation and denoising model. With the trained DNN model, we perform dereverberation and denoising for all 7138 sentences in the CHiME-2 training set, 409 sentences in the development set and 330 sentences in the test datasets, and resynthesize time-domain signals to construct new training and test datasets. No post-processing is used in this experiment. We then train ASR models using the new training set, and test the ASR models using the new test set. The baselines are ASR models trained and tested using original sentences including both clean and reverberant noisy sentences in the CHiME-2 corpus.

We use the Kaldi toolkit [28] to train two ASR systems, using original sentences and processed sentences, respectively. The first ASR system is a standard GMM-HMM based system using MFCC features with triphone three-state models. Speaker adaptive training [1] is performed during the training stage. The second is a hybrid ASR system, which uses alignments achieved from the GMM-HMM system and then trains DNNs with Mel-frequency filter bank features. This training scheme is motivated by [35], which achieves excellent performance on the CHiME-2 corpus. Sequence training [18] is also incorporated into this system. All the baseline systems are trained on multiple reverberant and noisy conditions.

We evaluate ASR performance in terms of word error rates (WERs). As shown in Fig. 15, for both GMM and hybrid ASR systems, the systems trained on processed sentences achieve lower WERs than those trained on original sentences across all SNR conditions. DNN based dereverberation and denoising considerably boost ASR performance in low SNRs, where the improvements are 11.7% (absolute) for the GMM system and 3.3% for the hybrid system in -6 dB SNR. The advantage gradually decreases as the SNR increases, partly because the performance decrement caused by reverberation and noise becomes smaller. On average, the improvements from original sentences are 9.5% for the GMM system and 2.0% for the hybrid system, demonstrating that our approach can be used as a front-end to improve ASR performance. We mention that these ASR experiments aim to show the promise of DNN based dereverberation and denoising rather than reach the state-of-the-art results (see [25]).

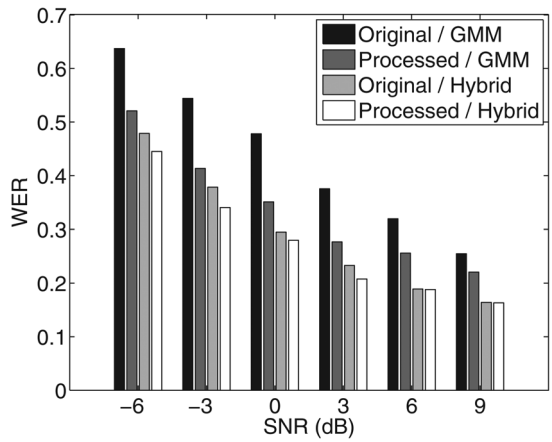

Fig. 15. ASR results. "Original / GMM" and "Processed / GMM" denote the results for the GMM-HMM systems using original sentences and processed sentences, respectively. "Original / Hybrid" and "Processed / Hybrid" denote the results for the hybrid systems using original sentences and processed sentences, respectively.

V. Discussion and Conclusion

We have proposed a supervised learning approach to perform dereverberation and denoising. A DNN is trained to learn a spectral mapping between corrupted speech and clean speech. Our approach extends that in [6], where pretraining is performed using a stack of restricted Boltzmann machines (RBMs) and a DNN is used to learn the spectral mapping in the cochleagram domain for dereverberation only. This study trains a DNN without RBM pretraining. Furthermore, the DNN here is trained to learn the spectrogram mapping to address dereverberation and denoising together. Although the cochleagram mapping slightly outperforms the spectrogram mapping for dereverberation in [6], it does not yield better results under reverberant and noisy conditions. To our knowledge, aside from pitch-based studies for separating reverberant voiced speech [30], [15], no study addresses both dereverebration and denoising through supervised learning. In addition, the present study employs iterative reconstruction to improve the DNN outputs and produces better resynthesized speech.

In this study, the feature is a concatenation of spectral features in a window, because temporal dynamics provides rich information for speech. A more fundamental approach to utilize temporal information would use a recurrent neural network (RNN), which is a natural extension of a feedforward network. An RNN aims to capture long-term temporal dynamics using time-delayed self-connections and is trained sequentially. We have trained RNN models for spectral mapping, and yielded around 0.2 dB improvement in terms of SNR_{fw}. Although this improvement is not significant, it is worth exploring RNNs in future work, for example, long short-term memory (LSTM) [9].

In our experiments, we train the DNN model using the IEEE corpus, which includes only one female speaker. In order to test generalizability, we have also conducted a cross-corpus experiment using the TIMIT corpus, where multiple speakers are contained in the test dataset. Our approach achieved similar

performance as that with the IEEE corpus. We have also conducted experiments where both training and test utterances are from TIMIT male speakers, and achieved the similar performance as female speakers. In addition, in our ASR experiments in Section IV-D using the CHiME-2 corpus training and testing were conducted in a speaker-independent manner, showing that our approach is robust to different speakers.

It is worth mentioning that we have trained a DNN based mapping on the cochleagram using the gammatone filterbank [38]. In this case, an element of an input vector corresponds to the log energy of each T-F unit of corrupted speech, while that of an output vector corresponds to the log energy of each T-F unit of clean speech. The DNN based cochleagram mapping also produces accurate cochleagram estimates, and the results are comparable with the spectrogram mapping.

In our ASR experiments, we resynthesize time-domain signals from DNN outputs and then perform speech recognition based on processed signals. According to our experiments, although the iterative signal reconstruction improves predicted speech intelligibility and quality scores, it does not lead to significant improvement for ASR performance. Comparing Fig. 2(c) with Fig. 2(e), the DNN output is still better than the spectrogram of the reconstructed signal, suggesting that we may extract MFCC or Mel filterbank features directly from the DNN output without resynthesis. As the DNN output is a better spectral representation than the spectrogram of resynthesized signals, it can be expected to yield better ASR performance. This should be explored in future work.

In summary, we have proposed to use DNNs to learn a spectral mapping from corrupted speech to clean speech for dereverberation, and dereverberation plus denoising. To our knowledge, this is the first study employing supervised learning to address the problem of speech dereverberation. Conceptually simple, our supervised learning approach significantly improves dereverberation, as well as denoising, performance in terms of predicted speech intelligibility and quality scores, and boosts ASR results in a range of reverberant and noisy conditions.

REFERENCES

[1] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, 1996, vol. 2, pp. 1137–1140.

[2] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *Proc. ICSLP*, 1996, vol. 2, pp. 889–892.

[3] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, 2011.

[4] D. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.

[5] E. Habets, Room Impulse Response Generator, 2010 [Online]. Available: http://home.tiscali.nl/ehabets/rirgenerator.html

[6] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. ICASSP*, 2014, pp. 4661–4665.

[7] O. Hazrati, J. Lee, and P. C. Loizou, "Blind binary masking for reverberation suppression in cochlear implants," *J. Acoust. Soc. Amer.*, vol. 133, pp. 1607–1614, 2013.

[8] 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.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] G. Hu, "Monaural speech organization and segregation," Ph.D. dissertation, The Ohio State Univ., Columbus, OH, USA, 2006.

[11] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, Sep. 2010.

[12] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, pp. 227–246, 1969.

[13] A. N. S. Institute, *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Melville, NY, USA: Acoust. Soc. of Amer., 1997.

[14] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 625–638, May 2009.

[15] Z. Jin, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1091–1102, Jul. 2011.

[16] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, pp. 1486–1494, 2009.

[17] B. Kingsbury and N. Morgan, "Recognizing reverberant speech with RASTA-PLP," in *Proc. ICASSP*, 1997, vol. 2, pp. 1259–1262.

[18] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proc. ICASSP*, 2009, pp. 3761–3764.

[19] K. Kokkinakis, O. Hazrati, and P. C. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *J. Acoust. Soc. Amer.*, vol. 129, pp. 3221–3232, 2011.

[20] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude stft spectrogram based on spectrogram consistency," in *Proc. Int. Conf. Digital Audio Effects*, 2010, vol. 10.

[21] K. Lebart, J.-M. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acust. united with Acust.*, vol. 87, no. 3, pp. 359–366, 2001.

[22] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3387–3405, 2009.

[23] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[24] A. K. Nabelek, "Communication in noisy and reverberant environments," in *Acoustical Factors Affecting Hearing Aid Performance*, G. A. Stubebaker and I. Hochberg, Eds., 2nd ed. ed. Needham Heights, MA, USA: Allyn and Bacon, 1992.

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

[26] *Speech Dereverberation*, P. Naylor and N. Gaubitch, Eds. New York, NY, USA: Springer, 2010.

[27] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, pp. 165–169, 1979.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011, pp. 1–4.

[29] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, vol. 2, pp. 749–752.

[30] N. Roman and D. L. Wang, "Pitch-based monaural segregation of reverberant speech," *J. Acoust. Soc. Amer.*, vol. 120, pp. 458–469, 2006.

[31] N. Roman and J. Woodruff, "Intelligibility of reverberant noisy speech with ideal binary masking," *J. Acoust. Soc. Amer.*, vol. 130, pp. 2153–2161, 2011.

[32] N. Roman, "Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold," *J. Acoust. Soc. Amer.*, vol. 133, no. 3, pp. 1707–1717, 2013.

[33] S. O. Sadjadi and J. H. L. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *Proc. ICASSP*, 2011, pp. 5448–5451.

[34] 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.

[35] Y. Tachioka, S. Watanabe, J. Le Roux, and J. R. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," in *Proc. CHiME'13*, 2013, pp. 19–24.
[36] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. ICASSP*, 2013, pp. 126–130.
[37] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, 2010.
[38] *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. L. Wang and G. J. Brown, Eds. Hoboken, NJ, USA: Wiley, 2006.
[39] Y. Wang and D. L. Wang, "Feature denoising for speech separation in unknown noisy environments," in *Proc. ICASSP*, 2013, pp. 7472–7476.
[40] M. Wu and D. L. Wang, "A one-microphone algorithm for reverberant speech enhancement," in *Proc. ICASSP*, 2003, pp. 844–847.
[41] M. Wu, "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.
[42] Y. Xu, J. Du, L. Dai, and C. 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.
[43] X. Zhao, Y. Wang, and D. L. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 836–845, Apr. 2014.
[44] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Commun.*, vol. 9, no. 4, pp. 351–356, 1990.

Kun Han received the Ph.D. degree from the Ohio State University in computer science in 2014 and the M.S. degree from the University of Science and Technology of China, Hefei, China, in 2008. Since 2014, he has been a Research Scientist at Facebook. His research interests include speech processing and machine learning.

Yuxuan Wang photograph and biography not available at the time of publication.

DeLiang Wang photograph and biography not available at the time of publication.

William S. Woods photograph and biography not available at the time of publication.

Ivo Merks (M'05) studied applied physics at Delft University of Technology, Delft, The Netherlands. He received his Masters of Science (ir.) degree in 1995 and his Ph.D. (dr.) degree in 2000. Since 2004, he has been with Starkey Hearing Technologies, where he is currently a Principal Research Engineer. His research interests include audio signal processing applications for hearing aids, such as noise reduction, beamforming, and feedback cancellation.

Tao Zhang received his B.S. degree in physics with a major in acoustics from Nanjing University, Nanjing, China, in 1986, his M.S. degree in electrical engineering from Peking University, Beijing, China, in 1989 and his Ph.D. degree in speech and hearing science from The Ohio State University, Columbus, OH, USA, in 1995. Since 2001, he has been with Starkey Hearing Technologies, where he is currently Director of Signal Processing Research. His research interests include signal processing research for hearing instruments, ultra-low power, real-time embedded systems, acoustics and psychoacoustics.