Two-Stage Deep Learning for Noisy-Reverberant Speech Enhancement

Yan Zhao, *Student Member, IEEE*, Zhong-Qiu Wang, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—In real-world situations, speech reaching our ears is commonly corrupted by both room reverberation and background noise. These distortions are detrimental to speech intelligibility and quality, and also pose a serious problem to many speech-related applications, including automatic speech and speaker recognition. In order to deal with the combined effects of noise and reverberation, we propose a two-stage strategy to enhance corrupted speech, where denoising and dereverberation are conducted sequentially using deep neural networks. In addition, we design a new objective function that incorporates clean phase during model training to better estimate spectral magnitudes, which would in turn yield better phase estimates when combined with iterative phase reconstruction. The two-stage model is then jointly trained to optimize the proposed objective function. Systematic evaluations and comparisons show that the proposed algorithm improves objective metrics of speech intelligibility and quality substantially, and significantly outperforms previous one-stage enhancement systems.**

***Index Terms*—Deep neural networks, denoising, dereverberation, phase, ideal ratio mask.**

I. Introduction

IN DAILY listening environments, reverberation from surface reflections in a room and background noise from other sound sources both distort target speech. These distortions, particularly when combined, can severely degrade speech intelligibility for human listeners, especially for hearing-impaired (HI) listeners [9]. Moreover, a lot of speech-related tasks, such as automatic speech recognition (ASR) and speaker identification (SID), become more difficult under these adverse conditions [1], [7], [23]. Therefore, solutions to denoising and dereverberation will benefit human listeners and many speech processing applications.

Manuscript received February 19, 2018; revised July 6, 2018 and September 5, 2018; accepted September 10, 2018. Date of publication September 17, 2018; date of current version October 15, 2018. This work was supported in part by the AFOSR under Grant FA9550-12-1-0130, in part by the NIH under Grant R01 DC012048, and in part by a Starkey research gift, and the Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tuomas Virtanen. *(Corresponding author: Yan Zhao.)*

Y. Zhao and Z.-Q. Wang are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: zhao.836@osu.edu; wangzhon@cse.ohio-state.edu).

D. 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, and also with the Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: dwang@cse.ohio-state.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.2018.2870725

Given the importance of the problem, a lot of effort has been made in the past decades to combat noise and reverberation [24], [27], [33]. In recent years, deep neural networks (DNNs) have been widely employed for speech enhancement or separation. Substantially better performance over conventional speech enhancement methods has been reported in many studies [16], [37], [39], [42]. The basic idea is to formulate the enhancement or separation problem as a supervised learning problem, and then utilize DNNs for supervised learning. For more discussion on deep learning based speech separation, we refer the interested reader to a recent comprehensive review [34].

Despite the large number of recent studies for DNN-based speech enhancement, few address both denoising and dereverberation. To separate noisy-reverberant speech, Han *et al.* [15] proposed a spectral mapping algorithm to perform denoising and dereverberation simultaneously using a single DNN. The idea is to learn a mapping function from the spectrum of noisy-reverberant speech to that of clean-anechoic speech. However, informal listening with human subjects indicates no evident improvement on speech intelligibility. Zhao *et al.* [44] pointed out that this is likely because different natures of noise and reverberation make them difficult to address together. Generally speaking, background noise is an additive signal to clean speech, while reverberation is a convolution process with a room impulse response (RIR) [43]. Taking this difference and human tolerance of room reverberation [8] into account, Zhao *et al.* [44] learned a mapping function to the spectrum of noise-free reverberant speech, without performing dereverberation. On this task, they reported speech intelligibility improvements for HI listeners in some noisy-reverberant conditions. Another recent work on noisy-reverberant speech enhancement is time-frequency masking in the complex domain by Williamson and Wang [40]. They introduced a complex ideal ratio mask (cIRM) using clean-anechoic speech as the desired signal for DNN-based enhancement. Experiment evaluations show that cIRM estimation outperforms Han *et al.*'s spectral mapping approach [15]. Like the spectral mapping method, they attempt to remove noise and room reverberation in one processing stage.

We believe that denoising and dereverberation should be addressed separately, due to their fundamental differences. In this paper, we propose a two-stage system to enhance noisy-reverberant speech. In the proposed system, we first develop two DNN-based subsystems that are trained for denoising and dereverberation individually. Then, we concatenate these pre-trained DNNs to perform joint training. It is worth noting that

the strategy of performing denoising and dereverberation in a step by step fashion was adopted previously [21]. This study first removes additive noise by using spectral subtraction, and then removes late reverberation from the noise-suppressed reverberant signal by a multi-step linear prediction dereverberation algorithm. Different from this previous study, we use DNNs for denoising and dereverberation and perform joint training afterwards. Previously, joint optimization has been applied to robust speech recognition tasks [6], [25].

Furthermore, motivated by the time-domain signal reconstruction technique [38], we propose a new objective function that incorporates clean phase to calculate the mean squared error (MSE) in the time domain. We find that this new objective function leads to consistently better performance in objective speech intelligibility and quality metrics.

The main contributions of this study are twofold. First, we propose a DNN-based two-stage framework to enhance noisy-reverberant speech. Second, we incorporate clean speech phase into the objective function to perform system optimization to obtain a better magnitude spectrum estimate. A preliminary version of our study is published in [45]. Compared with the previous conference paper, in this study, we investigate the two-stage system in more noisy-reverberant conditions, including more types of noise, untrained signal-to-noise ratios (SNRs), untrained reverberation times, recorded RIRs, and different speakers.

This paper is organized as follows. In Section II, we first describe the signal model for the noisy-reverberant speech. Then, the proposed two-stage enhancement algorithm and objective function are presented. Experimental setup and evaluation results are given in Section III. We conclude this paper in Section IV.

II. Algorithm Description

In this section, the noisy-reverberant signal model is first introduced. We then describe our noisy-reverberant speech enhancement algorithm. Fig. 1 shows the diagram of the proposed system, which consists of three modules: a denoising module, a dereverberation module and a time-domain signal reconstruction (TDR) module. We point out that the TDR module is only utilized during training and removed during testing. The three modules are introduced in detail in the following subsections.

A. Signal Model

Let $s(t)$, $x(t)$, $n(t)$ and $h(t)$ denote anechoic speech, reverberant speech, background noise and room impulse response function, respectively. The noisy-reverberant speech $y(t)$ can be modelled by

$$y(t) = x(t) + n(t) = s(t) * h(t) + n(t) \tag{1}$$

where $*$ stands for the convolution operator.

The objective of this study is to recover the anechoic signal $s(t)$ from the corresponding noisy-reverberant observation $y(t)$. This mathematical model suggests the proper order of denoising and dereverberation. Since $n(t)$ is generally uncorrelated with the desired signal $s(t)$ and the reverberant speech $x(t)$, it is natural to remove the noise from $y(t)$ first and then to recover the anechoic speech.

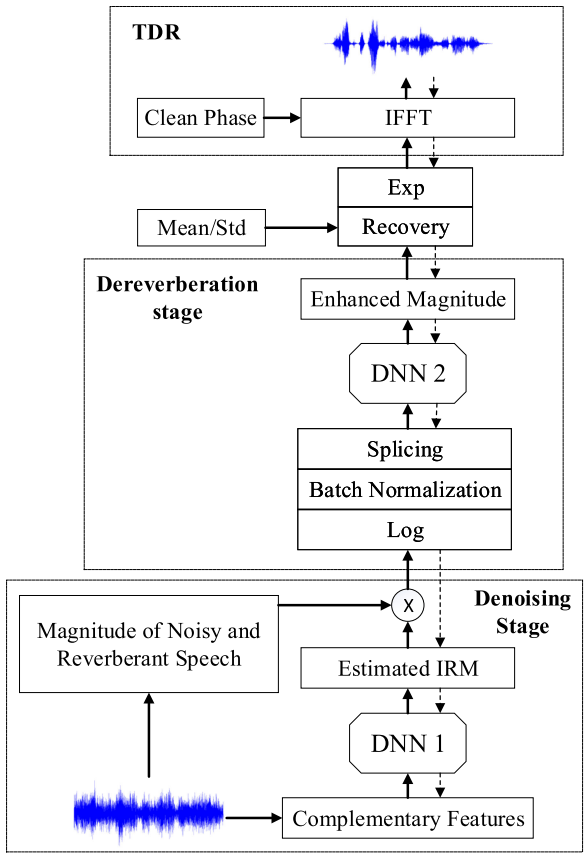


Fig. 1. System diagram of the proposed two-stage model, where "Std" denotes standard deviation and "Exp" exponential operation.

B. Denoising Stage

Given a noisy-reverberant utterance, the aim of this stage is to remove the background noise while keeping the reverberation untouched. In other words, the target signal of this processing stage is the noise-free reverberant speech. In order to suppress noise, we adopt the commonly used time-frequency (T-F) masking framework. Specifically, the ideal ratio mask (IRM) is estimated by employing DNN. Then, the predicted mask is applied to the T-F representation of noisy-reverberant speech to perform denoising as shown in Fig. 1. Recent research [3] using DNN to estimate the IRM for segregating speech from background noise has shown substantial speech intelligibility improvements for HI listeners.

Within DNN-based speech enhancement, an alternative method is to directly estimate the log-magnitude or log-power spectrum of clean speech [36], [42]. However, a study on training targets [36] suggests that masking-based targets tend to outperform mapping-based ones in both objective speech intelligibility and quality. With room reverberation added, our previous work [44] also indicates that the adoption of masking-based targets can bring significant performance improvements over mapping-based targets.

Based on the above observations, we employ a DNN with 3 hidden layers to predict the IRM in order to remove the noise from noisy-reverberant speech. Since the training target, the IRM, is bounded between 0 and 1, a sigmoid activation function is used in the output layer.

The denoising IRM is defined as follows [36],

$$IRM(m, f) = \sqrt{\frac{X^2(m, f)}{X^2(m, f) + N^2(m, f)}} \tag{2}$$

where $X^2(m, f)$ and $N^2(m, f)$ denote the energy of reverberant speech and background noise, respectively, at time frame m and frequency channel f. As shown in Fig. 1, the magnitude spectrum of noisy-reverberant speech is then multiplied by an estimated IRM to form the input to the next dereverberation stage processing.

A set of complementary features is adopted as the input for this stage [35], i.e., 15-dimensional amplitude modulation spectrogram (AMS), 31-dimensional Mel-frequency cepstral coefficients (MFCC), 13-dimensional relative spectral transform perceptual linear prediction (RASTA-PLP), 64-dimensional Gammatone filterbank power spectra (GF), and their deltas. Therefore, for each time frame, the feature dimension is 246 ($2 \times (15 + 31 + 13 + 64)$). We note that this set of features is originally introduced for denoising in anechoic environments. A 11-frame context window is utilized to encompass the input features (see Section III).

C. Dereverberation Stage

After the suppression of background noise, the original problem is reduced to recovering the anechoic speech $s(t)$ from the reverberant speech $x(t)$. To perform dereverberation in this stage, we follow the spectral mapping method proposed by Han *et al.* [14]. The choice of spectral mapping is also motivated by the fact that masking is well justified for separation as speech and noise are uncorrelated but uncorrelatedness does not hold well for dereverberation [34]. Compared with the original spectral mapping algorithm, our dereverberation stage has two major differences. First, instead of using percent normalization (normalizing values to the range of [0, 1]), we normalize the training target, log-magnitude spectrum of clean-anechoic speech, to zero mean and unit variance by using the global statistics of training data. This normalization is suggested in [41], which indicates that, in contrast to percent normalization, mean-variance normalization retains more spectral details, hence beneficial for the recovery of anechoic spectrum. Second, we use the IRM-processed magnitude spectrum of noisy-reverberant speech for feature extraction to train the dereverberation DNN. A log compression and mean-variance normalization are also applied to the features before splicing adjacent frames (11 frames in this study). By using IRM-processed features, we expect closer coupling between the separately trained denoising stage and dereverberation stage, which can benefit joint training. The DNN used in this stage has 3 hidden layers as well and a linear layer is used as the output layer.

D. Time-Domain Signal Reconstruction With Clean Phase

Most supervised speech separation systems perform enhancement on the magnitude spectrum and use the noisy phase to synthesize the time-domain signal. In order to alleviate the mismatch between the enhanced magnitude and the noisy phase, Wang and Wang [38] proposed a DNN to learn to perform TDR given the noisy phase. Improvements on objective speech quality are reported by their method. Similarly, Erdogan *et al.* [5] proposed to predict a phase-sensitive mask. Le Roux *et al.* [22] pointed out that the objective function of TDR with the noisy phase is equivalent to that of phase-sensitive masking. However, with the noisy-reverberant phase, Wang and Wang's approach could be problematic, since the phase is corrupted more seriously under such conditions. On the other hand, the magnitude and phase spectra carry complementary information [26], which implies that phase can be potentially utilized to help obtain better magnitude enhancement. Motivated by these observations, we extend the TDR method and propose a new objective function. More specifically, during training, we feed the enhanced magnitude (after denoising and dereverberation) to an inverse fast Fourier transform (IFFT) layer to reconstruct the enhanced time-domain signal with the corresponding clean phase, and then optimize the loss in the time domain. While phase-sensitive masking also utilizes clean phase in the form of the phase difference between clean speech and corrupted speech, our proposed method directly employs clean phase in training.

Mathematically, at time frame m, let $\boldsymbol{s}$, $\hat{\boldsymbol{S}}$ and $\boldsymbol{p}_c$ denote the windowed clean-anechoic signal segment, the corresponding enhanced magnitude after two-stage processing and clean phase, respectively. Θ denotes the parameters of a learning system. Then, the objective function at the training stage is defined as follows,

$$\mathcal{L}(\boldsymbol{s}, \hat{\boldsymbol{S}}; \Theta) = \|\boldsymbol{s} - IFFT(\hat{\boldsymbol{S}} \circ e^{j\boldsymbol{p}_c})\|_2^2 \tag{3}$$

where $\circ$ denotes the element-wise multiplication and $\|\cdot\|_2$ the L_2 norm.

From another perspective, supervised speech enhancement systems typically consider all the T-F units to be of the same importance and ignore the underlying energy of the corrupted or desired signal in each T-F unit. In the proposed objective function, computing the loss in the time domain will force the learning machine to implicitly place more emphasis on the T-F units that contribute more to the time-domain signal. In other words, instead of weighting T-F units explicitly using normalized mixture energy [19] or mixture energy [39], our method weights different units on the basis of their corresponding time-domain signal.

E. Joint Training

As shown in Fig. 1, we concatenate the denoising DNN and the dereverberation DNN into a bigger network for joint optimization. In the denoising stage, an estimated IRM is applied to the magnitude spectrogram of noisy-reverberant speech. The enhanced magnitude is then passed through a log function to compress the dynamic range of values. We add a

batch normalization layer [17] before the splicing operation to make sure that the input to the dereverberation DNN is properly normalized. During training, this layer keeps exponentially moving averages on the mean and standard deviation of each mini-batch. During testing, such running mean and standard deviation are fixed to perform normalization. The normalized features of 11 frames (see Section III) are spliced as the input features to the dereverberation DNN. After the dereverberation stage, the enhanced log-magnitude is recovered by using the standard deviation and mean of clean-anechoic log-magnitude, as we have normalized the target of dereverberation DNN before training. These statistics are computed from the training data. Finally, after an exponential operation, the processed magnitude is fed to the IFFT layer to get the enhanced time-domain signal. The loss in the time domain is computed by (3). Since each step above is differentiable, we can derive the error gradients to jointly train the whole system.

Before joint training, the denoising DNN and the dereverberation DNN are trained separately with the MSE objective function, and the resulting parameters are used to initialize the two-stage speech enhancement system.

F. Time-Domain Signal Resynthesis

During system training, clean phase is used to help obtain better magnitude spectrum. However, such information is not available during testing, and the observed phase may be severely distorted by room reverberation and background noise. In order to reduce the inconsistency between the enhanced magnitude and the corrupted phase, Griffin-Lim's iterative phase enhancement algorithm [12] and overlap-add (OLA) method are employed to resynthesize the enhanced time-domain signal (see also [15]). In other words, during testing, the TDR module (namely, the IFFT layer with clean phase) is removed, and the output of the proposed system is the enhanced magnitude. Since the accuracy of estimated phase of Griffin-Lim's method depends on the accuracy of the enhanced spectral magnitude [10], with more accurate magnitude estimation, our system is expected to perform better enhancement.

III. Evaluations and Comparisons

A. Datasets and Experimental Setup

We evaluate our proposed system on the IEEE corpus [29] spoken by a female speaker. There are 72 phonetically balanced lists in this corpus and each list contains 10 sentences. Sentences selected from list 1–50, list 68–72 and list 58–67 are utilized to construct training data, validation data and test data, respectively. Therefore, the sentences in each set are different from those in the others. The RIRs are generated in a simulated room, whose size is 10 m $\times$ 7 m $\times$ 3 m. We generate different RIRs with the position of the receiver fixed and the position of the speaker randomly chosen. Moreover, we keep the distance between the receiver and the speaker to 1 m, so that the direct to reverberant ratio (DRR) does not change much under each T_{60}. We utilize an RIR generator [13] to produce the RIRs, which is based on the image method [2]. In the experiments, three values of T_{60} are investigated, i.e., 0.3 s, 0.6 s and 0.9 s. For each T_{60}, 10 RIRs are generated for the training and validation sets; 1 RIR is generated for the test set. To further evaluate the generalization of our approach under different values of T_{60}, for each value of T_{60} at 0.4 s, 0.5 s, 0.7 s and 0.8 s, we also generate 1 RIR for testing. In summary, we have 500 $\times$ 3 (T_{60}s) $\times$ 10 (RIRs) = 15,000 reverberant utterances in the training set, 50 $\times$ 3 (T_{60}s) $\times$ 10 (RIRs) = 1,500 reverberant utterances in the validation set, and 100 $\times$ 7 (T_{60}s) $\times$ 1 (RIR) = 700 reverberant utterances in the test set. The average DRR values are listed in Table I.

TABLE I
Average DRR Values at Different Reverberation Time

T_{60} (s)	0.3	0.4	0.5	0.6	0.7	0.8	0.9
DRR (dB)	4.96	2.95	1.69	0.73	0.18	-0.12	-0.99

Four types of noises are studied, including speech shaped noise (SSN), babble noise (BABBLE), noise recorded in a living room (DLIVING), and cafeteria noise recorded in a busy office cafeteria (PCAFETER), with SSN being stationary and the others nonstationary. The DLIVING noise and PCAFETER noise are from the Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [32]. Since we are dealing with monaural noisy-reverberant speech enhancement, the first channel recorded DLIVING and PCAFETER are used in the experiments. It should be noted that, since DLIVING and PCAFETER are recorded in real rooms, they contain room reverberation. All the noises are about 5 min long. To generate noisy-reverberant speech, random cuts from the first 4 min and the remaining 1 min of each noise are mixed with the reverberant speech at a specified SNR for the training/validation set and test set, respectively. For training and validation, three levels of SNRs are used, namely, -6 dB, 0 dB and 6 dB, where reverberant speech is taken as the signal in calculating the SNR. For testing, besides the three SNRs seen during training, -3 dB and 3 dB are also included to evaluate the generalization of our system to mismatched SNR levels. Consequently, for each type of noise, there are 15,000 $\times$ 3 (SNRs) = 45,000 utterances for training, 1,500 $\times$ 3 (SNRs) = 4,500 utterances for validation, and 700 $\times$ 5 (SNRs) = 3,500 utterances for testing. DNN models are trained and evaluated for each noise separately. However, neither the noise segments nor the RIRs of test data are seen during training.

In our experiments, given a signal sampled at 16 kHz, we divide the signal using a 20-ms Hamming window with a 10-ms window shift for framing. To optimize the time-domain loss, the clean-anechoic target signal of each frame is also windowed by using a Hamming window. For each time frame, a 320-point fast Fourier transform (FFT) is applied, resulting in 161 frequency bins. In order to incorporate the temporal information of adjacent frames, we utilize a context window to slide 5 frames on each side of the current frame (11 frames in total) to combine the frame-level information. This size of context window is suggested in [15]. To resynthesize time-domain signals, the number of iterations of Griffin-Lim's algorithm is set to 20.

For DNN training, the input features are normalized to zero mean and unit variance by using the statistics of the training data. All DNNs are trained with exponential linear units (ELUs) [4], which lead to faster convergence and better performance over rectified linear units (ReLUs) [11], especially when networks become deep. In each hidden layer, there are 1024 hidden units. We utilize Adam [20] as the optimizer to train the networks. Dropout regularization [30] is adopted to prevent overfitting. The dropout rates for the input layer and all the hidden layers are set to 0.2. The hyper-parameters are chosen according to the performance on the validation data. Since the IRM target is bounded by [0, 1], we employ sigmoid activation units in the output layer; for the others, linear output layers are used.

B. Comparison Methods

We compare our proposed system to two DNN-based speech enhancement methods. One is the spectral mapping method [15]. A DNN is trained to directly estimate the log-magnitude of clean-anechoic speech from the log-magnitude of noisy-reverberant speech. Like the dereverberation stage described in Section II-C, we normalize the target log-magnitude of clean-anechoic speech to zero mean and unit variance. For convenience, we denote this method as "**mapping**". The other approach is denoted as "**masking**". It is an extension of the IRM masking method for the noisy speech enhancement [36]. To enhance the noisy-reverberant speech, we construct the IRM by taking the clean-anechoic speech as desired signal and the rest as interference. A DNN with the complementary features as the input is utilized to predict the thus defined IRM, and then the estimated ratio mask is applied to the magnitude of noisy-reverberant speech to obtain the enhanced signal. In order to maintain the same network depth with our proposed two-stage system, for these two comparison systems, we employ a DNN with 6 hidden layers, each with the same size as in the proposed system.

Our proposed objective function in the time domain utilizes clean phase for training, and it can be employed by other supervised speech enhancement systems. In order to examine the impact of the proposed objective function, we change the objective function of the comparison methods to the proposed one, and denote these two new systems as "**mapping+TDR**" and "**masking+TDR**", with the latter network structure similar to that proposed in [38]. Note that these two DNNs are initialized by using the parameters of the corresponding mapping method and masking method. In addition, the phase-sensitive mask (PSM) [5] is evaluated as an objective function, and it is compared with the proposed objective function. Specifically, a DNN with the complementary features as the input is used to estimate the PSM defined in [5]. The DNN has the same network architecture as used in the masking comparison method. This method is denoted as "**PSM**".

The proposed two-stage system is denoted as "**two-stage+TDR**". In order to investigate how much performance change is due to the two-stage strategy alone, another two-stage system without the TDR module is also included as another comparison. This method is denoted as "**two-stage**".

C. Evaluation Metrics

In the experiments, we evaluate speech intelligibility by using short-time objective intelligibility (STOI) [31], which predicts speech intelligibility by comparing the temporal envelops of the clean reference speech and the processed speech. The value range of STOI is typically from 0 to 1. In addition, perceptual evaluation of speech quality (PESQ) [28] is employed to evaluate speech quality. PESQ scores are in the range $[-0.5, 4.5]$. For both metrics, the higher scores indicate better performance. Since our study intends to remove both background noise and room reverberation, clean-anechoic speech is used as the reference signal to compute the objective metrics.

D. Evaluation Results

1) One-Stage Processing vs. Two-Stage Processing: Tables II, III, IV and V list the STOI and PESQ scores of unprocessed and processed signals under different noisy-reverberant conditions by using different methods. We first consider the matched conditions; in other words, the SNRs and reverberation times are seen during training. Similar performance trends were observed for the four different types of noise. By switching to the IRM target defined for removing noise and reverberation and using the complementary features, the masking method outperforms the mapping method at each noisy-reverberant condition in terms of both STOI and PESQ metrics. This observation is consistent with the denoising results reported in [36], [44].

Taking the masking method as the stronger one-stage processing baseline, we compare its average performance with that of the two-stage method. In terms of STOI, the two-stage strategy brings 0.9%, 1.0%, 3.4% and 1.8% improvements over the masking method for DLIVING, PCAFETER, BABBLE and SSN, respectively. Note that, for some noisy-reverberant conditions, the masking method is slightly better than the two-stage method. For example, under the condition of $T_{60} = 0.3$ s and SNR $= 0$ dB with DLIVING, the STOI value of the masking method is 0.7% higher than that of the two-stage method. Such cases happen only when T_{60} is low. For some noises, with low reverberation time and high DRR, the noisy-reverberant speech enhancement problem is to an extent reduced to the noisy speech enhancement problem. In other words, in the two-stage framework, the denoising stage becomes dominant. Since we employ a smaller DNN for the denoising stage than that used in the masking method, the performance of the two-stage method can be a little worse. In terms of PESQ, performance improvements over the masking method are observed in most conditions.

2) Effects of Proposed Objective Function: To study the effects of the proposed objective function, we combine it with the mapping method and the masking method, respectively. Clearly, additional performance gain is obtained by employing the new objective function. On average, for DLIVING, the mapping method and the masking method are improved by 2.4% and 1.7% STOI values, respectively; for PCAFETER, they are improved by 2.4% and 1.6%, respectively; for BABBLE, they are improved by 3.3% and 1.8%, respectively; and for SSN, the STOI values are increased by 3.2% and 1.5%, respectively. At the same time, there are also some improvements in PESQ

TABLE II
STOI AND PESQ SCORES IN MATCHED CONDITIONS FOR DLIVING NOISE. BOLDFACE HIGHLIGHTS THE BEST RESULT OF EACH CONDITION

	STOI (in %)										**PESQ**									
T_{60} (s)	**0.3**			**0.6**			**0.9**			**Avg.**	**0.3**			**0.6**			**0.9**			**Avg.**
SNR (dB)	**-6**	**0**	**6**	**-6**	**0**	**6**	**-6**	**0**	**6**		**-6**	**0**	**6**	**-6**	**0**	**6**	**-6**	**0**	**6**	
unprocessed	75.5	84.1	89.8	71.3	78.0	82.4	64.2	71.6	75.5	76.9	1.36	1.77	2.12	1.36	1.68	1.89	1.08	1.41	1.56	1.58
mapping	82.0	88.4	90.3	80.2	87.1	88.9	73.7	83.0	85.7	84.4	1.99	2.36	2.50	1.89	2.25	2.37	1.56	1.99	2.10	2.11
mapping+TDR	85.3	90.3	92.1	83.8	89.0	90.7	78.0	84.9	87.4	86.8	2.16	2.47	2.65	2.04	2.34	2.48	1.72	2.07	2.18	2.23
masking	86.8	91.9	94.3	83.7	88.2	90.0	78.9	85.2	87.6	87.4	2.23	2.59	2.85	2.17	2.46	2.61	1.82	2.17	2.30	2.36
masking+TDR	88.0	**92.8**	**95.0**	85.8	90.0	91.7	81.3	87.3	89.8	89.1	2.34	2.69	2.94	2.25	2.52	2.68	1.93	2.25	2.40	**2.46**
PSM	86.8	92.0	94.4	83.8	88.6	90.8	78.6	85.4	87.9	87.6	2.33	**2.72**	**3.00**	2.26	**2.56**	**2.73**	1.90	2.26	2.39	**2.46**
two-stage	86.6	91.2	93.0	85.4	90.0	91.7	80.4	86.8	89.2	88.3	2.27	2.58	2.75	2.18	2.46	2.60	1.90	2.25	2.39	2.38
two-stage+TDR	**88.4**	92.5	94.2	**87.2**	**91.1**	**92.7**	**82.8**	**88.0**	**90.5**	**89.7**	**2.37**	2.66	2.85	**2.27**	2.52	2.67	**1.98**	**2.30**	**2.45**	2.45

TABLE III
STOI AND PESQ SCORES IN MATCHED CONDITIONS FOR PCAFETER NOISE

	STOI (in %)										**PESQ**									
T_{60} (s)	**0.3**			**0.6**			**0.9**			**Avg.**	**0.3**			**0.6**			**0.9**			**Avg.**
SNR (dB)	**-6**	**0**	**6**	**-6**	**0**	**6**	**-6**	**0**	**6**		**-6**	**0**	**6**	**-6**	**0**	**6**	**-6**	**0**	**6**	
unprocessed	59.9	73.0	83.0	56.0	68.4	77.6	50.9	61.9	70.4	66.8	1.10	1.39	1.74	1.05	1.38	1.68	0.98	1.20	1.42	1.33
mapping	65.9	81.3	87.6	60.9	78.5	85.8	53.8	73.3	81.7	74.3	1.21	1.86	2.25	1.04	1.72	2.15	0.91	1.53	1.89	1.62
mapping+TDR	69.0	83.0	89.2	64.5	80.6	87.7	57.3	75.6	83.6	76.7	1.37	1.96	2.36	1.23	1.84	2.25	1.04	1.63	1.97	1.74
masking	70.8	84.6	91.0	65.8	80.6	87.2	58.5	75.5	83.7	77.5	1.43	2.00	2.43	1.33	1.92	2.33	1.12	1.69	2.04	1.81
masking+TDR	72.0	**85.5**	**91.9**	67.4	82.3	88.9	60.0	77.6	85.9	79.1	1.50	2.08	**2.55**	1.36	1.97	2.42	1.16	1.77	2.15	1.88
PSM	70.2	84.7	91.2	65.0	80.8	87.9	57.2	75.4	84.0	77.4	1.46	2.07	2.54	1.35	1.97	2.42	1.10	1.73	2.12	1.86
two-stage	71.0	84.4	90.1	66.6	82.6	89.1	59.5	77.7	85.7	78.5	1.48	2.08	2.46	1.29	1.95	2.37	1.12	1.76	2.15	1.85
two-stage+TDR	**71.9**	85.2	91.2	**68.1**	**83.3**	**90.0**	**61.1**	**78.9**	**86.9**	**79.6**	**1.54**	**2.11**	2.53	**1.40**	**2.01**	**2.43**	**1.21**	**1.83**	**2.22**	**1.92**

TABLE IV
STOI AND PESQ SCORES IN MATCHED CONDITIONS FOR BABBLE NOISE

	STOI (in %)										**PESQ**									
T_{60} (s)	**0.3**			**0.6**			**0.9**			**Avg.**	**0.3**			**0.6**			**0.9**			**Avg.**
SNR (dB)	**-6**	**0**	**6**	**-6**	**0**	**6**	**-6**	**0**	**6**		**-6**	**0**	**6**	**-6**	**0**	**6**	**-6**	**0**	**6**	
unprocessed	52.5	66.8	79.4	50.4	62.3	73.7	47.0	57.3	66.9	61.8	1.07	1.22	1.55	1.01	1.22	1.52	0.96	1.12	1.31	1.22
mapping	66.2	81.4	86.9	63.8	78.2	84.5	55.2	71.8	79.8	74.2	1.24	1.95	2.28	1.16	1.80	2.11	0.99	1.53	1.87	1.66
mapping+TDR	70.2	83.8	89.0	68.5	81.0	86.8	60.1	75.4	82.6	77.5	1.40	2.05	2.39	1.33	1.91	2.23	1.11	1.63	1.96	1.78
masking	70.9	84.2	90.5	68.1	80.5	86.5	60.6	75.4	82.9	77.7	1.42	2.01	2.38	1.40	1.94	2.29	1.15	1.68	2.01	1.81
masking+TDR	72.9	85.4	91.5	70.6	82.4	88.2	62.1	77.6	85.2	79.5	1.56	2.15	2.55	1.52	2.04	2.39	1.27	1.81	2.13	1.94
PSM	69.8	83.8	90.5	66.4	80.1	86.7	59.0	74.8	82.9	77.1	1.43	2.07	2.48	1.43	2.00	2.37	1.16	1.72	2.07	1.86
two-stage	75.1	86.4	90.7	73.5	84.3	89.1	65.2	79.8	86.0	81.1	1.68	2.24	2.56	1.61	2.12	2.41	1.37	1.90	2.20	2.01
two-stage+TDR	**77.2**	**87.1**	**91.6**	**76.0**	**85.2**	**90.0**	**68.4**	**81.1**	**87.1**	**82.6**	**1.79**	**2.28**	**2.59**	**1.75**	**2.17**	**2.45**	**1.50**	**1.96**	**2.24**	**2.08**

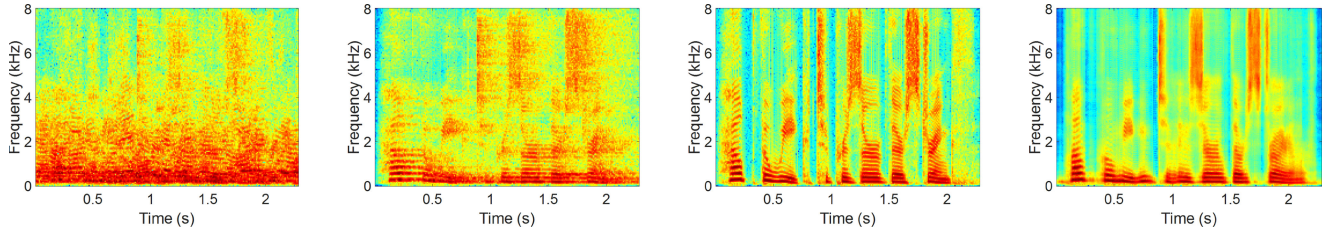
TABLE V
STOI AND PESQ SCORES IN MATCHED CONDITIONS FOR SSN

	STOI (in %)										**PESQ**									
T_{60} (s)	**0.3**			**0.6**			**0.9**			**Avg.**	**0.3**			**0.6**			**0.9**			**Avg.**
SNR (dB)	**-6**	**0**	**6**	**-6**	**0**	**6**	**-6**	**0**	**6**		**-6**	**0**	**6**	**-6**	**0**	**6**	**-6**	**0**	**6**	
unprocessed	55.2	68.7	81.4	53.6	64.7	75.0	50.2	59.6	68.9	64.1	0.71	1.06	1.48	0.78	1.09	1.47	0.69	1.01	1.26	1.06
mapping	68.4	83.2	88.3	66.2	80.8	86.3	59.1	74.3	82.1	76.5	1.41	2.05	2.36	1.33	1.91	2.21	1.14	1.63	1.97	1.78
mapping+TDR	72.8	85.6	90.3	70.7	83.7	88.4	63.9	77.7	84.4	79.7	1.54	2.14	2.46	1.48	2.04	2.32	1.27	1.75	2.02	1.89
masking	74.0	85.9	91.4	71.9	82.7	87.4	65.6	78.2	84.6	80.2	1.58	2.11	2.47	1.56	2.05	2.35	1.31	1.77	2.08	1.92
masking+TDR	75.2	87.1	**92.5**	73.2	84.7	89.2	66.8	80.1	86.8	81.7	1.68	2.21	2.58	1.63	2.13	2.43	1.39	1.88	2.18	2.01
PSM	71.8	85.0	91.2	69.5	81.8	87.4	63.1	76.8	84.4	79.0	1.61	2.16	2.55	1.60	2.09	2.43	1.33	1.82	2.14	1.97
two-stage	75.5	87.6	91.5	73.7	85.6	89.8	66.4	80.9	86.8	82.0	1.69	2.31	2.59	1.63	2.18	2.46	1.38	1.93	2.22	2.04
two-stage+TDR	**77.1**	**88.6**	92.4	**75.6**	**86.9**	**90.7**	**68.9**	**82.2**	**87.8**	**83.4**	**1.77**	**2.34**	**2.62**	**1.71**	**2.24**	**2.49**	**1.48**	**1.99**	**2.26**	**2.10**

values after incorporating the new objective function. These results suggest that the proposed objective function provides an effective way to improve supervised speech enhancement in general.

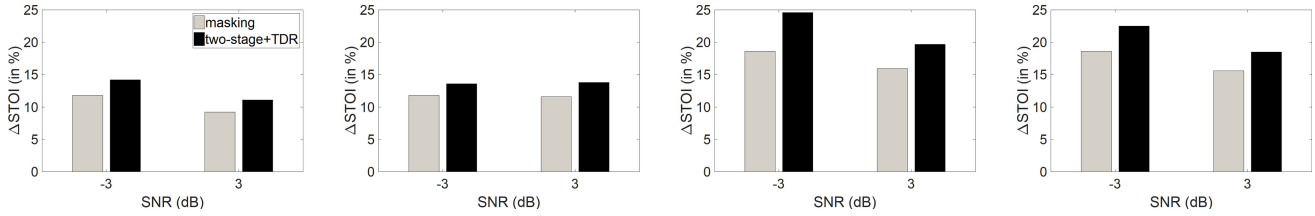
The introduction of a phase-sensitive objective function in the PSM provides higher PESQ improvements than ratio masking, suggesting better speech quality. The possible reason is that, compared to the IRM, the phase-sensitive mask produces higher signal-to-distortion ratio [5]. However, with the proposed objective function, the masking+TDR method outperforms the PSM method in most conditions. This also demonstrates the effectiveness of the proposed objective function.

With this new objective function and the two-stage processing strategy, our final model, two-stage+TDR, performs



(a) noisy-reverberant speech (b) reverberant speech (c) anechoic speech (d) enhanced speech

Fig. 2. (Color online) Spectrograms of a noisy-reverberant utterance (BABBLE noise, SNR $= -6$ dB, $T_{60} = 0.9$ s), a reverberant utterance ($T_{60} = 0.9$ s), an anechoic utterance and an enhanced utterance (two-stage+TDR). The sentence is "Help the weak to preserve their strength."



(a) DLIVING noise (b) PCAFETER noise (c) BABBLE noise (d) SSN

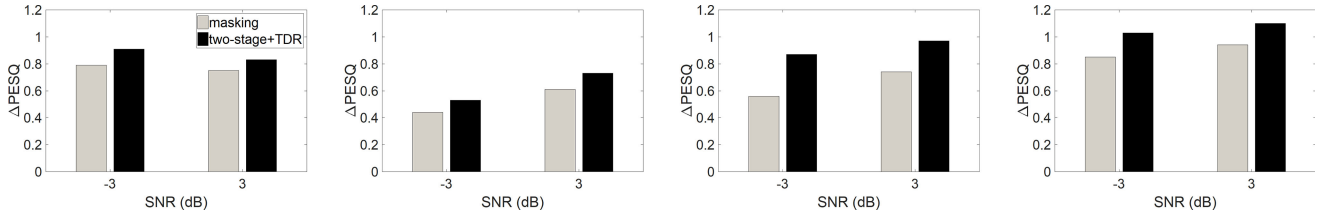
Fig. 3. ΔSTOI evaluations at untrained SNRs for 4 types of noise.

the best in the large majority of noisy-reverberant conditions. Compared with the masking method, on average, 2.3%, 2.1%, 4.9% and 3.2% STOI improvements are obtained for DLIVING, PCAFETER, BABBLE and SSN, respectively. We should point out that the masking method is a very strong baseline for comparison. By comparing with Han *et al.*'s spectral mapping method, we can get a larger performance boost. To illustrate the proposed algorithm, an enhancement example is presented in Fig. 2. The spectrogram of a noisy-reverberant utterance with BABBLE noise at SNR $= -6$ dB and reverberation time of 0.9 s is shown in Fig. 2(a). The corresponding spectrograms of reverberant speech, anechoic speech and speech enhanced by the proposed algorithm (two-stage+TDR) are presented in Figs. 2(b), (c) and (d), respectively. Comparing the spectrograms of noisy-reverberant speech and enhanced speech, one can see that smearing effects resulting from reverberation and additive noise have been largely removed, and the spectrotemporal structure is considerably restored. This indicates that the proposed system performs denoising and dereverberation effectively.

3) Evaluations at Untrained SNRs: For supervised approaches, the generalization ability of a trained model is a critical factor to consider. Here, we evaluate how well our proposed method generalizes to untrained conditions during testing. In the following evaluations, the STOI score change (ΔSTOI in percent) and the PESQ score change (ΔPESQ) are used as the criterion. We compare the proposed algorithm with the strong baseline, i.e., the masking method. To evaluate the generalization to untrained SNR conditions, under each of the two untrained SNR levels (-3 dB and 3 dB), we take average across the reverberation times (0.3 s, 0.6 s and 0.9 s). The results are presented in Fig. 3 and Fig. 4. As shown in the figures, for all untrained SNR conditions with different types of noise, our method improves objective speech intelligibility and quality substantially, and outperforms the masking method. This indicates that, even though the SNR levels are not included in the training data, the proposed approach generalizes well to these untrained SNR conditions.

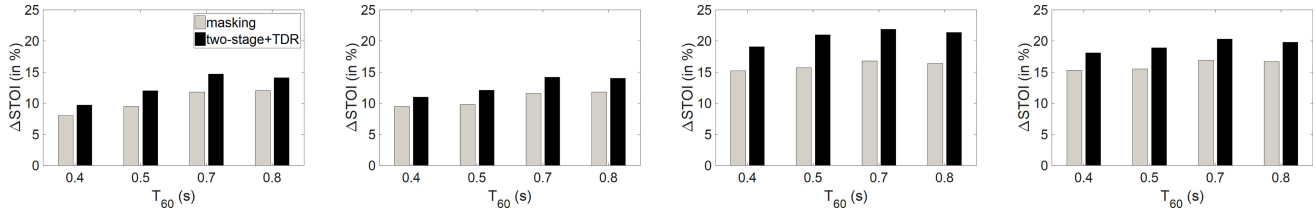
4) Evaluations at Untrained Reverberation Times: The model is trained at three reverberation times, namely, 0.3 s, 0.6 s and 0.9 s. We now test the model under new reverberation times to evaluate its generalization ability to a wide range of reverberation times. In this evaluation, we use reverberation times of 0.4 s, 0.5 s, 0.7 s and 0.8 s. For presentation, average ΔSTOI score and ΔPESQ score across different SNR levels (-6 dB, 0 dB and 6 dB) at each reverberation time are computed. The evaluation results are shown in Fig. 5 and Fig. 6. Similar to the evaluation of the untrained SNRs, at each untrained reverberation time, objective speech intelligibility and quality scores are improved. The proposed algorithm obtains a consistent performance improvement over the baseline masking method.

5) Evaluations With Recorded RIRs: Simulated RIRs are used to generate reverberant speech in the above experiments. Now we evaluate our trained systems with recorded RIRs. Three RIRs are selected from the Aachen Impulse Response (AIR) database [18], and resampled to 16 kHz. They were recorded in a lecture room, a meeting room and an office room, and the corresponding T_{60} is 0.70 s, 0.21 s and 0.37 s, respectively. With sentences from list 58–67 of the IEEE corpus, these recorded RIRs and untrained segments of the 4 types of noise, we construct a new test set for evaluation. Specifically, under each noise, there are 1,500 utterances, i.e. 100 (sentences) $\times$ 3 (RIRs) $\times$ 5 (SNRs, -6, -3, 0, 3, 6 dB). In each reverberant room condition, we take average of ΔSTOI and ΔPESQ scores across the 5 SNRs. The comparison results with the masking method are given in Fig. 7 and Fig. 8 in ΔSTOI and ΔPESQ. Although the enhancement systems are trained with very limited reverberant conditions (e.g. three reverberation times, fixed room size, fixed speaker-microphone distance and uniformed reflection surfaces), they generalize well to



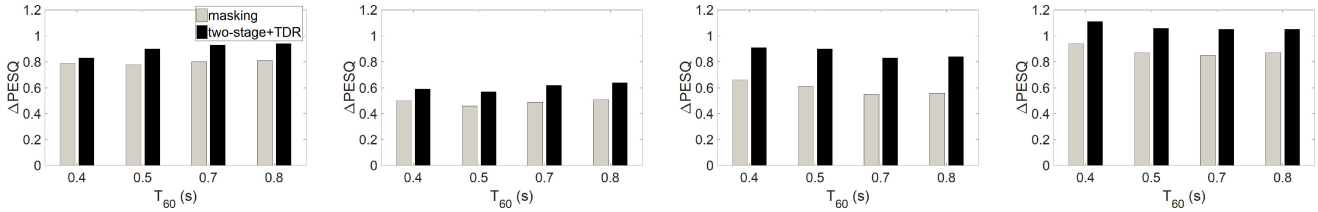
(a) DLIVING noise (b) PCAFETER noise (c) BABBLE noise (d) SSN

Fig. 4. ΔPESQ evaluations at untrained SNRs for 4 types of noise.



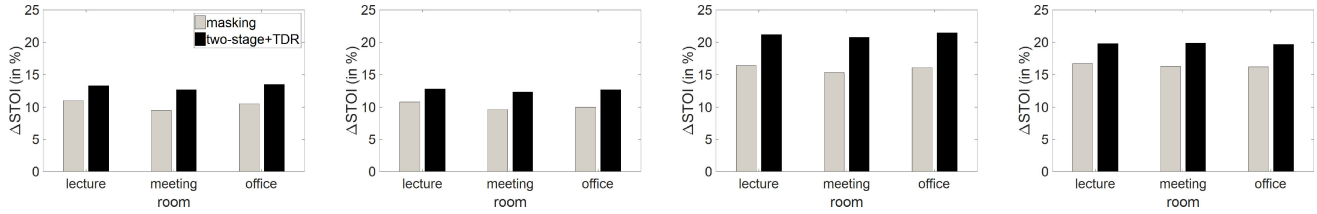
(a) DLIVING noise (b) PCAFETER noise (c) BABBLE noise (d) SSN

Fig. 5. ΔSTOI evaluations at untrained reverberation times for 4 types of noise.



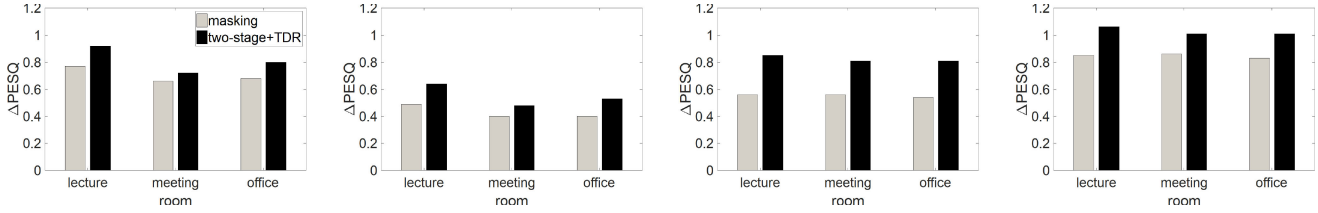
(a) DLIVING noise (b) PCAFETER noise (c) BABBLE noise (d) SSN

Fig. 6. ΔPESQ evaluations at untrained reverberation times for 4 types of noise.



(a) DLIVING noise (b) PCAFETER noise (c) BABBLE noise (d) SSN

Fig. 7. ΔSTOI evaluations at recorded RIRs for 4 types of noise.



(a) DLIVING noise (b) PCAFETER noise (c) BABBLE noise (d) SSN

Fig. 8. ΔPESQ evaluations at recorded RIRs for 4 types of noise.

real reverberant rooms. Substantial improvements in objective speech intelligibility and quality are obtained over unprocessed noisy-reverberant speech. Like previous evaluations, the proposed two-stage system outperforms the strong masking baseline.

6) Evaluations on Different Speakers: To illustrate that DNN models are not very sensitive to training speakers, we now evaluate the methods on speech utterances from different speakers. We emphasize that the models trained on a single speaker are used without any change or retraining in this evaluation. Specifically, 10 female speakers are randomly selected from the TIMIT corpus [46] and 2 sentences are selected from each speaker. Noise and room reverberation are added to construct a new noisy-reverberant test set. Therefore, under each

TABLE VI
STOI AND PESQ SCORES IN MISMATCHED SPEAKER CONDITIONS FOR DIFFERENT TYPES OF NOISE. BOLDFACE HIGHLIGHTS THE BEST RESULT. DLI, PCA AND BAB STAND FOR DLIVING, PCAFETER AND BABBLE, RESPECTIVELY

	STOI (in %)				PESQ			
	DLI	PCA	BAB	SSN	DLI	PCA	BAB	SSN
unprocessed	71.6	61.7	56.7	59.2	1.90	1.60	1.46	1.37
mapping	71.5	60.6	57.5	67.7	1.57	1.23	1.14	1.45
mapping+TDR	75.1	64.1	62.1	70.6	1.85	1.47	1.37	1.64
masking	78.8	66.6	66.7	72.6	**2.28**	**1.79**	**1.72**	1.89
masking+TDR	**79.5**	67.4	68.0	73.4	2.19	1.70	1.64	1.81
PSM	79.0	66.4	66.1	72.6	2.26	1.75	**1.72**	**1.92**
two-stage	76.9	66.2	67.9	73.2	1.86	1.47	1.48	1.67
two-stage+TDR	79.4	**68.5**	**70.7**	**74.6**	2.05	1.66	1.66	1.78

noise, there are 180 utterances, i.e., 10 (speakers) $\times$ 2 (sentences) $\times$ 3 (RIRs for $T_{60} = 0.3$, 0.6, 0.9 s) $\times$ 3 (SNRs, -6, 0, 6 dB). We report average STOI and PESQ scores across the three SNRs and the three T_{60}s. The evaluation results are listed in Table VI. Compared with the average performance shown in Tables II–V in the matched speaker condition, the improvements over unprocessed noisy-reverberant speech become smaller as expected. But Table VI still indicates that the proposed two-stage+TDR model trained with just one speaker shows some generalization to different speakers, and outperforms the other systems in term of STOI. It is worth noting that the mapping method seems unable to generalize in this case and performs rather poorly in PESQ. This could be a reason why the two-stage+TDR system does not provide best PESQ scores as the second stage performs spectral mapping.

IV. CONCLUDING REMARKS

Background noise and room reverberation are two major causes distorting speech signal in real listening environments. To address both the noise and reverberation problem, we have proposed a two-stage algorithm to deal with two kinds of distortions in sequence. Two DNNs are first utilized to perform denoising and dereverberation separately, and then combined into a deeper network subject to joint optimization with a new proposed objective function in the time domain. By incorporating clean phase, the system can be directly optimized in the time domain, leading to a better estimate of magnitude spectrum. We also show that the proposed objective function can be directly adopted by other supervised separation systems, and boosts their performance. Systematic evaluations using objective speech intelligibility and quality metrics show that our system outperforms a spectral mapping method and a stronger masking-based method in various noisy-reverberant conditions. These evaluation results strongly indicate that the proposed system improves actual speech intelligibility and quality in real noisy-reverberant environments.

During the evaluation, we trained separate models for different types of noise. We have also carried out multi-noise training for further evaluation. In other words, we trained the DNN models on all four noises. Compared with single-noise training, slightly lower STOI and PESQ scores were obtained under some low SNR conditions; for most conditions, however, similar scores were observed to noise-specific training. In addition, the same performance trends were observed and the proposed method still performed best.

In this study, we have developed a framework for joint optimization of two enhancement subsystems. For the denoising subsystem and dereverberation subsystem, we utilize relatively straightforward DNNs. A recent study [41] proposed a reverberation-time-aware DNN (RTA-DNN) to perform dereverberation, which takes the reverberation time into account and outperforms a conventional DNN-based method. The RTA-DNN can be considered in our dereverberation stage. Moreover, although our proposed objective function employs clean phase during training, the aim is still to obtain a better estimate of magnitude spectrum. Since phase is corrupted more severely under noisy-reverberant conditions, if clean-anechoic phase can be partly estimated, better speech intelligibility and quality of the enhanced signal can be expected. Therefore, incorporating complex ratio masking [40] or extending the two-stage algorithm to the complex domain represent promising future directions.

REFERENCES

[1] K. A. Al-Karawi, A. H. Al-Noori, F. F. Li, and T. Ritchings, "Automatic speaker recognition system in adverse conditions—implication of noise and reverberation on system performance," *Int. J. Inf. Electron. Eng.*, vol. 5, pp. 423–427, 2015.

[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoustical Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[3] J. Chen, Y. Wang, S. E. Yoho, D. L. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoustical Soc. Amer.*, vol. 139, pp. 2604–2612, 2016.

[4] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. Int. Conf. Learn. Represent.*, 2016.

[5] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 708–712.

[6] T. Gao, J. Du, Y. Xu, C. Liu, L.-R. Dai, and C.-H. Lee, "Joint training of DNNs by incorporating an explicit dereverberation structure for distant speech recognition," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, p. 86, 2016.

[7] D. Gelbart and N. Morgan, "Double the trouble: Handling noise and reverberation in far-field automatic speech recognition," in *Proc. INTERSPEECH*, 2002.

[8] S. Gelfand and I. Hochberg, "Binaural and monaural speech discrimination under reverberation," *Audiology*, vol. 15, pp. 72–84, 1976.

[9] E. L. J. George, S. T. Goverts, J. M. Festen, and T. Houtgast, "Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners," *J. Speech, Lang., Hearing Res.*, vol. 53, pp. 1429–1439, 2010.

[10] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, Mar. 2015.

[11] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[12] D. W. 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.

[13] E. Habets. Room impulse response generator, 2014. [Online]. Available: https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator. Accessed on: 2017.

[14] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 4628–4632.

[15] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.
[16] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.
[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Machine Learn.*, 2015, pp. 448–456.
[18] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. IEEE Int. Conf. Digit. Signal Process.*, 2009, pp. 1–5.
[19] 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.
[20] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015.
[21] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Multi-step linear prediction based speech dereverberation in noisy reverberant environment," in *Proc. INTERSPEECH*, 2007, pp. 854–857.
[22] J. Le Roux, E. Vincent, and H. Erdogan, "Learning based approaches to speech enhancement and separation," in *Proc. INTERSPEECH Tutorials*, 2016.
[23] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.
[24] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.
[25] M. Mimura, S. Sakai, and T. Kawahara, "Joint optimization of denoising autoencoder and DNN acoustic model based on multi-target learning for noisy speech recognition," in *Proc. INTERSPEECH*, 2016, pp. 3803–3807.
[26] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Commun.*, vol. 81, pp. 1–29, 2016.
[27] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.
[28] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 2, pp. 749–752.
[29] E. H. Rothauser *et al.*, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, pp. 225–246, 1969.
[30] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.
[31] 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.
[32] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoustical Soc. Amer.*, vol. 133, pp. 3591–3591, 2013.
[33] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.
[34] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.
[35] Y. Wang, K. Han, and D. L. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 270–279, Feb. 2013.
[36] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[37] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.
[38] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4390–4394.
[39] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2014, pp. 577–581.
[40] D. S. Williamson and D. L. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, Jul. 2017.
[41] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 102–111, Jan. 2017.
[42] 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.
[43] T. Yoshioka *et al.*, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, Nov. 2012.
[44] Y. Zhao, D. L. Wang, I. Merks, and T. Zhang, "DNN-based enhancement of noisy and reverberant speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 6525–6529.
[45] Y. Zhao, Z.-Q. Wang, and D. L. Wang, "A two-stage algorithm for noisy and reverberant speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5580–5584.
[46] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Commun.*, vol. 9, pp. 351–356, 1990.

Yan Zhao (S'16) received the B.E. degree in information engineering from Xi'an Jiaotong University, Xi'an, China, in 2009, the M.S. degree in electrical and computer engineering from The Ohio State University, Columbus, OH, USA, in 2014, and the second M.S. degree in computer science and engineering from The Ohio State University, in 2018. He is currently working toward the Ph.D. degree with the Department of Computer Science and Engineering, The Ohio State University. His research interests include speech separation and machine learning.

Zhong-Qiu Wang, photograph and biography not available at the time of publication.

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