Deep Learning Based Target Cancellation for Speech Dereverberation

Zhong-Qiu Wang and DeLiang Wang, *Fellow, IEEE*

***Abstract*—This article investigates deep learning based single- and multi-channel speech dereverberation. For single-channel processing, we extend magnitude-domain masking and mapping based dereverberation to complex-domain mapping, where deep neural networks (DNNs) are trained to predict the real and imaginary (RI) components of the direct-path signal from reverberant (and noisy) ones. For multi-channel processing, we first compute a minimum variance distortionless response (MVDR) beamformer to cancel the direct-path signal, and then feed the RI components of the cancelled signal, which is expected to be a filtered version of non-target signals, as additional features to perform dereverberation. Trained on a large dataset of simulated room impulse responses, our models show excellent speech dereverberation and recognition performance on the test set of the REVERB challenge, consistently better than single- and multi-channel weighted prediction error (WPE) algorithms.**

***Index Terms*—Complex spectral mapping, phase estimation, microphone array processing, speech dereverberation, deep learning.**

I. Introduction

ROOM reverberation is pervasive in modern hands-free speech communication, such as teleconferencing and smart speakers. In a reverberant enclosure, speech signals propagate in the air and are inevitably reflected by the walls, ceiling, floor, and any objects in the room. As a result, the signal captured by a distant microphone is a summation of an infinite number of delayed and decayed copies of original source signals. Room reverberation is known to be detrimental to automatic speech recognition (ASR) systems, and severely degrades speech quality and intelligibility. Speech dereverberation is a challenging task, as reverberation is a convolutive interference, unlike background noise which is additive, and it is difficult to distinguish the direct-path signal from its reverberated versions, especially when room reverberation is strong or environmental noise is also present [1].

Manuscript received September 18, 2019; revised January 13, 2020 and February 5, 2020; accepted February 6, 2020. Date of publication February 28, 2020; date of current version March 13, 2020. This work was supported in part by the NIDCD under Grant R01 DC012048, in part by the NSF under Grant ECCS-1808932, and in part by Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Timo Gerkmann. *(Corresponding author: Zhong-Qiu Wang.)*

Zhong-Qiu Wang is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: wangzhon@cse.ohio-state.edu).

DeLiang Wang is with the Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASLP.2020.2975902

For single-channel dereverberation, one conventional approach estimates the power spectral density (PSD) of late reverberation to compute a Wiener-like filter based on, for example, an exponentially decaying model computed using an estimated reverberation time [2] and a relative convolutive transfer function model [3]. The weighted prediction error (WPE) algorithm [4], [5] is probably the most widely used algorithm for speech dereverberation. It uses variance-normalized delayed linear prediction to predict late reverberation from past observations, and subtracts the predicted reverberation to estimate target speech. It iteratively estimates the time-varying PSD of target speech and the linear filter, and is unsupervised in nature. Many ASR studies report that WPE suppresses reverberation with low speech distortions, and consistently improves ASR performance even for multi-conditionally trained ASR backends [6].

When multiple microphones are available, spatial information can be leveraged to filter out signals not arriving from the estimated target direction. Single-channel WPE can be extended to multi-channel WPE [4] by simply concatenating the observations across multiple microphones when performing linear prediction. Another popular approach for multi-channel speech dereverberation is the so-called suppression approach [7], [8], which decomposes a multi-channel Wiener filter into a product of a time-invariant or time-varying MVDR beamformer and a monaural Wiener post-filter. This approach requires accurate estimation of spatial covariance matrices and PSDs. It can utilize the phase produced by linear beamforming, which is expected to be better than the mixture phase, since MVDR beamforming is distortionless. However, the phase improvement is dependent on linear beamforming, which is less effective when room reverberation is strong or when the number of microphones is small. In addition, the Wiener post-filter is a real-valued mask, and would inevitably introduce phase inconsistency problems [9], [10], when directly applied to the beamformed signal for enhancement.

Different from conventional algorithms, supervised learning based approaches train a DNN to predict the magnitudes or real-valued masks of the direct-path signal from reverberant observations [11]–[15]. Such data-driven approaches typically lead to good magnitude (or PSD) estimation compared with conventional algorithms [1], thanks to the non-linear modeling power of DNN. However, the DNN operates in the magnitude domain, and mixture phase is typically utilized for signal resynthesis. Phase estimation is hence a promising direction for further improvement. Another direction in dereverberation uses DNN estimated speech magnitudes as the PSD estimate for WPE

[16]–[19]. This approach can leverage the spectral structure in speech for linear prediction, and most importantly eliminates the iterative process, making WPE suitable for online processing. In offline scenarios, although ASR improvement is observed on the eight-channel task of the REVERB challenge, it leads to slightly worse performance on the single-channel task [16].

In this context, our study extends magnitude-domain masking and mapping based speech dereverberation to the complex domain, where a DNN is trained to predict the RI components of direct sound from reverberant ones. Although previous studies perform single-channel complex masking or mapping for speech denoising [20]–[22], their results in reverberant conditions are mixed [23] and how to extend to multi-channel processing is unclear.

Our study approaches multi-channel dereverberation from the angle of target cancellation, where a key assumption is that the target speaker is a directional source, and is typically non-moving within a short utterance. This suggests that we can point a null beam to cancel the target speaker, and the remaining signal would only contain a filtered version of reverberation. This filtered reverberation can be utilized as extra features for DNN to perform multi-channel complex spectral mapping based dereverberation. It should be noted that similar ideas of target cancellation were explored in binaural speech segregation [24] and multi-channel dereverberation [25], [8]. Their purposes are, however, different (e.g., on the PSD estimation of late reverberation), and they do not address phase estimation.

Our study makes four main contributions. First, we extend deep learning based magnitude-domain single-channel speech dereverberation to the complex domain for phase estimation. The phase estimation method follows the complex spectral mapping idea in [21], [22], while our study addresses direct sound vs. reverberation and noise, rather than speech vs. noise in anechoic conditions. Second, we introduce for complex spectral mapping a magnitude-domain loss function, which dramatically improves speech quality measures in reverberant conditions. Third, we propose a novel target cancellation strategy to utilize spatial information to improve the estimation of direct sound. Fourth, we investigate the effectiveness of DNN based phase estimation for beamforming and post-filtering, while the DNN in previous deep learning based multi-channel enhancement operates in the magnitude domain. We emphasize that the proposed algorithms are designed in a way such that the resulting models, once trained, can be directly applied to arrays with an arbitrary number of microphones arranged in an unknown geometry.

The rest of this paper is organized as follows. We introduce the physical model in Section II. The proposed algorithms are detailed in Section III, followed by experimental setup in Section IV. Evaluation and comparison results are presented in Section V. Conclusions are made in Section VI.

II. Physical Model and Objectives

Given a P-microphone time-domain signal $\boldsymbol{y}[n] = [y_1[n], \ldots, y_P[n]]^{\mathrm{T}} \in \mathbb{R}^{P\times 1}$ recorded in a reverberant and noisy enclosure, the physical model in the short-time Fourier transform (STFT) domain is formulated as:

$$\begin{aligned}\boldsymbol{Y}(t,f) &= \boldsymbol{c}(f;q)\, S_q(t,f) + \boldsymbol{H}(t,f) + \boldsymbol{N}(t,f) \\ &= \boldsymbol{S}(t,f) + \boldsymbol{V}(t,f) \end{aligned} \quad (1)$$

where $S_q(t,f) \in \mathbb{C}$ are the complex STFT values of the direct-path signal of the target speaker captured by a reference microphone q at time t and frequency f, $\boldsymbol{c}(f;q) \in \mathbb{C}^{P\times 1}$ is the relative transfer function with the q^{th} element being one, and $\boldsymbol{c}(f;q)S_q(t,f)$, $\boldsymbol{H}(t,f)$, $\boldsymbol{N}(t,f)$ and $\boldsymbol{Y}(t,f) \in \mathbb{C}^{P\times 1}$ respectively represent the STFT vectors of the direct-path signal, reverberation, reverberant noise and received mixture at a T-F unit.

We propose multiple deep learning algorithms to enhance the mixture Y_q capture at the reference microphone q to recover S_q, by exploiting single- and multi-channel information contained in $\boldsymbol{Y}$. In this study, $\boldsymbol{N}(t,f)$ is assumed to be a quasi-stationary air-conditioning noise, as our focus is on dereverberation; the proposed algorithms can be straightforwardly applied to deal with more noises. The target speaker is assumed to be still within an utterance. Our study also assumes offline scenarios: we normalize the time-domain sample variance of each input multi-channel signal $\boldsymbol{y}$ to one before any processing. This normalization would be important for mapping-based enhancement to deal with random gains in input signals.

In the remainder of this paper, we refer to $\boldsymbol{S}(t,f) = \boldsymbol{c}(f;q)S_q(t,f)$ as the target component we aim to extract, and $\boldsymbol{V}(t,f) = \boldsymbol{H}(t,f) + \boldsymbol{N}(t,f)$ as the non-target component to remove.

III. Proposed Algorithms

There are two DNNs in our system. The first DNN performs single-channel dereverberation by predicting the RI components of the direct-path signal from a mixture. Dereverberation results are utilized to compute an MVDR beamformer. The second DNN utilizes the RI components of beamformed speech as additional features to further improve the estimation of the RI components of the direct-path signal. Fig. 1 illustrates the overall system.

A. Single-Channel Complex Spectral Mapping

Following recent studies [21], [22], we train a DNN to directly predict the RI components of the direct sound from reverberant and noisy ones. One key difference is that [21] and [22] deal with speech vs. noise, while our study addresses direct sound vs. reverberation and noise. We use the following loss function

$$\mathcal{L}_{\mathrm{RI}} = \|\hat{R}_p - \mathrm{Real}(S_p)\|_1 + \|\hat{I}_p - \mathrm{Imag}(S_p)\|_1, \quad (2)$$

where $p \in \{1, \ldots, P\}$ indexes microphones, $\hat{R}_p$ and $\hat{I}_p$ are the estimated RI components obtained by using linear activation in the output layer, and $\mathrm{Real}(\cdot)$ and $\mathrm{Imag}(\cdot)$ respectively extract the RI components. The enhanced speech at microphone p is obtained as $\hat{S}_p^{(k)} = \hat{R}_p^{(k)} + j\hat{I}_p^{(k)}$, where the superscript $k \in \{1, 2\}$ denotes the output from the k^{th} DNN, as shown in Fig. 1.

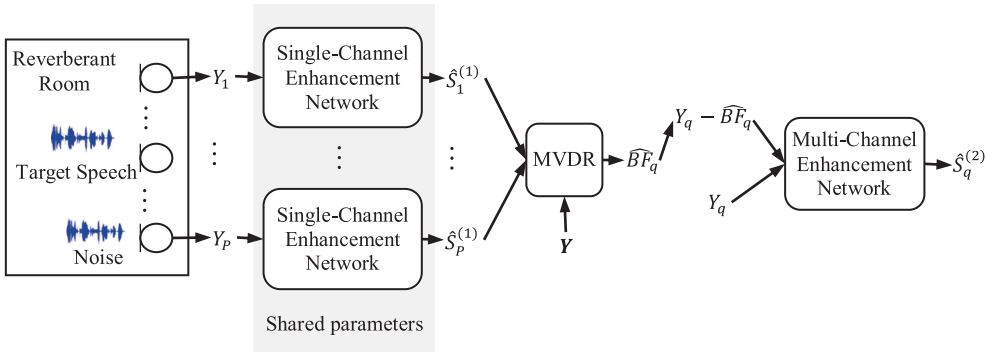

Fig. 1. Illustration of overall system for single- and multi-channel speech dereverberation (or enhancement). There are two DNNs, one for single-channel and the other for multi-channel dereverberation and denoising. The superscript in $\hat{S}_1^{(1)}, \ldots, \hat{S}_P^{(1)}$ and $\hat{S}_q^{(2)}$ denotes the DNN used for processing.

Following recent studies that combine $\mathcal{L}_{\text{RI}}$ with a magnitude-domain loss [21], [26], we design the following loss function

$$\mathcal{L}_{\text{RI+Mag}} = \mathcal{L}_{\text{RI}} + \left\| \sqrt{\hat{R}_p^2 + \hat{I}_p^2} - |S_p| \right\|_1 \tag{3}$$

Different from [21], [26], our study does not compress the estimated magnitudes or complex spectra using logarithmic or power compression. This way, the DNN is always trained to estimate a complex spectrum that has consistent magnitude and phase structures, and therefore would likely produce a consistent estimated STFT spectrum at run time [10].

Our experiments show that including a loss on magnitude leads to large improvements in objective measures of speech quality, along with a small degradation on time-domain signal-to-noise ratio (SNR) based measures, compared with only using $\mathcal{L}_{\text{RI}}$.

B. Multi-Channel Complex Spectral Mapping

We propose a target cancellation approach to exploit spatial information for dereverberation. The motivation is that given an oracle MVDR beamformer $\boldsymbol{w}(f;q)$, the beamformed signal is distortion-less, meaning that $S_q(t,f) = \boldsymbol{w}(f;q)^H \boldsymbol{S}(t,f)$. Therefore, the difference between the mixture and the beamformed signal at reference microphone q, computed as

$$\begin{aligned} Y_q(t,f) - BF_q(t,f) &= Y_q(t,f) - \boldsymbol{w}(f;q)^H \boldsymbol{Y}(t,f) \\ &= S_q(t,f) + V_q(t,f) - \Big(\boldsymbol{w}(f;q)^H \boldsymbol{S}(t,f) \\ &\quad + \boldsymbol{w}(f;q)^H \boldsymbol{V}(t,f) \Big) \\ &= V_q(t,f) - \boldsymbol{w}(f;q)^H \boldsymbol{V}(t,f) \end{aligned} \tag{4}$$

would only contain a filtered version of non-target signals, i.e., $V_q(t,f) - \boldsymbol{w}(f;q)^H \boldsymbol{V}(t,f)$. Intuitively, the more microphones there are and the more accurate the beamformer is, the weaker the beamformed non-target speech $\boldsymbol{w}(f;q)^H \boldsymbol{V}(t,f)$ would be, and the closer $V_q(t,f) - \boldsymbol{w}(f;q)^H \boldsymbol{V}(t,f)$ is to the actual non-target speech $V_q(t,f)$ we aim to remove at microphone q. This makes $Y_q - \widehat{BF}_q$ a highly discriminative feature for dereverberation, and hence motivates us to use it as an extra input for DNN to predict S_q via complex spectral mapping. Without this feature, the DNN may struggle at distinguishing direct-path signal from its reverberated versions, as the latter is a summation of the delayed and decayed copies of the former.

We apply the single-channel complex spectral mapping model to each microphone signal and directly use the estimated speech $\hat{\boldsymbol{S}}^{(1)}$ to robustly compute an MVDR beamformer for cancelling target speech. Our study considers time-invariant MVDR (TI-MVDR) beamforming, as the target speaker is assumed still within each utterance, and reverberation and the considered noise are largely diffuse. The covariance matrices are computed as

$$\begin{aligned} \hat{\Phi}^{(s)}(f) &= \frac{1}{T} \sum_t \hat{\boldsymbol{S}}(t,f)\ \hat{\boldsymbol{S}}(t,f)^H \\ \hat{\Phi}^{(v)}(f) &= \frac{1}{T} \sum_t \hat{\boldsymbol{V}}(t,f)\, \hat{\boldsymbol{V}}(t,f)^H \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{V}}(t,f) = \boldsymbol{Y}(t,f) - \hat{\boldsymbol{S}}(t,f)$. The motivation is that the estimated complex spectra are expected to have cleaner phase than the mixture phase. In contrast, mask-weighted ways of computing covariance matrices (see Eq. (10) for example) [27]–[31] are fundamentally limited when there are insufficient T-F units dominated by the direct-path signal, such as when room reverberation or environmental noise is very strong.

The relative transfer function is then computed in the following way

$$\hat{\boldsymbol{r}}(f) = \mathcal{P}\left\{ \hat{\Phi}^{(s)}(f) \right\} \tag{6}$$

$$\hat{\boldsymbol{c}}(f;q) = \frac{\hat{\boldsymbol{r}}(f)}{\hat{r}_q(f)} \tag{7}$$

where $\mathcal{P}\{\cdot\}$ extracts the principal eigenvector. The motivation is that $\hat{\Phi}^{(s)}(f)$ would be close to a rank-one matrix if accurately estimated. Its principal eigenvector is therefore a reasonable estimate of the steering vector [32]. We then use Eq. (7) to obtain an estimated transfer function relative to a reference microphone q. We emphasize that, without using Eq. (7), a different complex gain would be introduced at each frequency, leading to distortions in the beamformed signal.

A TI-MVDR beamformer is then computed as

$$\hat{\boldsymbol{w}}(f;q) = \frac{\hat{\Phi}^{(v)}(f)^{-1}\hat{\boldsymbol{c}}(f;q)}{\hat{\boldsymbol{c}}(f;q)^{H}\hat{\Phi}^{(v)}(f)^{-1}\hat{\boldsymbol{c}}(f;q)} \tag{8}$$

The beamformed signal is obtained using

$$\widehat{BF}_q(t,f) = \hat{\boldsymbol{w}}(f;q)^{H}\boldsymbol{Y}(t,f) \tag{9}$$

For multi-channel dereverberation, we feed the RI components of $Y_q - \widehat{BF}_q$**,** in addition to the RI components of Y_q, to a DNN to estimate the RI components of the direct-path signal S_q (see Fig. 1).

We point out that this strategy is in spirit similar to the classic generalized sidelobe canceller (GSC) [32], which contains three components: a delay-and-sum (DAS) beamformer computed to enhance the target signal, a blocking matrix used to block the target signal, and an adaptive noise canceller designed to cancel the sidelobes produced by the DAS beamformer based on the blocked signal. The key difference here is that we compute an MVDR beamformer to block the target signal, and use deep learning to cancel the non-target signal in Y_q based on $Y_q - \widehat{BF}_q$.

From the spatial feature perspective, popular for deep learning based multi-channel speech enhancement [33]–[36] and speaker separation [37], the RI components of $\widehat{BF}_q$ or $Y_q - \widehat{BF}_q$ can be considered as complex-domain spatial features, which can be utilized by the DNN to extract a target speech signal with specific spectral structure and arriving from a particular direction. Such features are more general than those previously proposed for improving magnitude estimation, such as plain interchannel phase differences (IPD) [38], cosine and sine IPD [39], and target direction compensated IPD and the magnitudes of beamformed mixtures [37].

IV. Experimental Setup

Our models for dereverberation are trained on reverberant and noisy data created by using simulated room impulse responses (RIRs) and recorded noises. We first measure the performance on a relatively matched simulated test set, and then evaluate the trained models directly on the test set of the REVERB challenge [40] to show their generalization ability. This section describes the datasets and the setup for model training, and several baseline systems for comparison purposes.

A. Datasets and Evaluation Setup

Following the REVERB challenge [40], our training data for dereverberation is generated using the WSJCAM0 corpus. Different from REVERB, which only uses 24 measured eight-channel RIRs to generate its training data, we use a much larger set of RIRs (in total 39,305 eight-channel RIRs for training) generated by an RIR generator[1] to simulate room reverberation. See Algorithm 1 for the detailed simulation procedure. For each utterance, we randomly generate a room with different room characteristics, speaker and microphone locations, microphone array characteristics, and noise levels. Our study considers eight-channel circular arrays with radius ranging from 3 to 10 cm. The target speaker is placed on the same plane as the array, at a distance randomly drawn from 0.75 to 2.5 m. The reverberation time (T60) is randomly sampled between 0.2 and 1.3 s. We use the training and test noise (mostly diffuse quasi-stationary fan noise) in REVERB to simulate noisy reverberant mixtures in our training and test sets, respectively. The SNR between the

[1][Online]. Available: https://github.com/ehabets/RIR-Generator.

Algorithm 1: Data Spatialization Process (Simulated RIRs).

Input: WSJCAM0;
Output: spatialized reverberant (and noisy) WSJCAM0;
REP[train] $= 5$; *REP[validation]* $= 4$; *REP[test]* $= 3$;
For *dataset* in {*train*, *validation*, *test}* set of WSJCAM0 **do**
  For each anechoic speech signal s in *dataset* **do**
    Repeat *REP[dataset]* times **do**
      - Sample room length r_x and width r_y from [5,10] m;
      - Sample room height r_z from [3, 4] m;
      - Sample mic array height a_z from [1, 2] m;
      - Sample array displacement n_x and n_y from $[-0.5,\ 0.5]$ m;
      - Place array center at $\langle \frac{r_x}{2} + n_x, \frac{r_y}{2} + n_y, a_z \rangle$ m;
      - Sample array radius a_r from [0.03, 0.1] m;
      - Sample angle of first mic angle ϑ from $[0, \frac{\pi}{4}]$;
      For $p = 1 : P(= 8)$ **do**
      - Place mic p at $\langle \frac{r_x}{2} + n_x + a_r \cos(\vartheta + (p-1)\frac{\pi}{4}), \frac{r_y}{2} + n_y + a_r \sin(\vartheta + (p-1)\frac{\pi}{4}), a_z \rangle$ m;
      End
      - Sample target speaker locations in the $0 - 360^{\circ}$ plane:
      $$\langle s_x, s_y, s_z(= a_z) \rangle$$
      such that the distance from target speaker to array center is in between [0.75, 2.5] m, and target speaker is at least 0.5 m from each wall;
      - Sample T60 from [0.2, 1.3] s;
      - Generate multi-channel impulse responses using RIR generator and convolve them with s;
      If *dataset* in {*train*, *validation*} do
      - Sample a P-channel noise signal n from the training noise of REVERB corpus;
      Else
      - Sample a P-channel noise signal n from the testing noise of REVERB corpus;
      End
      - Concatenate channels of reverberated s and n respectively, scale them to an SNR randomly sampled from [5, 25] dB, and add them to obtain reverberant and noisy mixture;
    End
  End
End

direct sound and reverberant noise of each mixture is randomly drawn between 5 and 25 dB. The average direct-to-reverberation energy ratio[2] (DRR) is -3.7 dB with 4.4 dB standard deviation. There are 39, 305 (7,861 $\times$ 5, $\sim$80 h), 2,968 (742 $\times$ 4, $\sim$6 h), and 3,264 (1,088 $\times$ 3, $\sim$7 h) eight-channel utterances in the training, validation and test set, respectively. Note that the training and the test speakers are different. We denote this test set as **Test Set I**. At run time, we randomly choose a subset of microphones for each test mixture for evaluation. This setup therefore covers a wide range of microphone geometry. We use the direct-path signal at a reference microphone (i.e., the signal corresponding to S_q) as the reference for metric computation, and the first microphone is always considered as the reference. For P-channel processing, we randomly select $P-1$ microphones from the non-reference microphones and always report the performance on the reference microphone. This way, we can directly compare single- and multi-channel processing as they are both evaluated using the same reference signals.

We apply the trained models, without re-training, to the test set of the REVERB corpus, which contains simulated as well as recorded reverberant and noisy mixtures. We first evaluate the enhancement performance of the trained models on the simulated test set (denoted as **Test Set II**), where six measured eight-channel RIRs are used to simulate 2,176 reverberant and noisy mixtures. The six RIRs are measured in small-, medium- and large-size rooms, where the T60s are 0.25, 0.5 and 0.7 s respectively, and the speaker to microphone distance is around 0.5 m in the near-field case and 2.0 m in the far-field case. Recorded environmental noise is added at an SNR of 20 dB. In the REVERB challenge setup, only the sample at n_d, which is the index corresponding to the highest value in the measured RIR, is used to compute the direct-path signal (i.e., reference signal) for metric computation. However, due to measurement inaccuracy, this may not be realistic, since the samples in a small window around n_d are typically considered as in the direct-path RIR [41]. A short segment of an example RIR from REVERB is shown in Fig. 3(a), where T60 is around 0.7 s. If we only use the sample at n_d to simulate the direct-path signal, the resulting DRR would be unrealistically low, as the samples around the peak exhibit non-negligible energy; as a result, the reverberation generated by the surrounding samples would be difficult to remove. These surrounding samples should be considered when computing the direct-path signal, as they are in a measured RIR. Also, the sound source may not be a point source strictly and for a 16 kHz sampling rate, one discrete sample can have around $340/16,000$ m measurement error, where 340 (m/s) is the sound speed in the air. Furthermore, simulated direct-path RIRs are usually computed based on low-pass filtering, and they will be similar to a Sinc function even for a point source [42]. In Fig. 3(b) we show an example direct-path RIR simulated using the RIR generator by setting the T60 parameter to zero. In our study, we hence use the samples in the range $[n_d - 0.0025 \times 16,000,\ n_d + 0.0025 \times 16,000]$ (i.e., a 5-ms window surrounding the peak) of the measured RIRs to compute the direct-path signal for metric computation. This strategy aligns with the setup in the ACE challenge [41]. We then evaluate the dereverberation models on the ASR task of REVERB (denoted as **REVERB ASR**). The test utterances are real recordings with T60 (reverberation time) around 0.7 s and the speaker to microphone distances approximately 1 m in the near-field case and 2.5 m in the far-field case. Both Test Set II and REVERB ASR use an eight-microphone circular array with a 20 cm diameter, and the target speaker is non-moving within each utterance. We follow a *plug-and-play* approach for ASR, where enhanced signals are directly fed into a multi-conditionally trained ASR backend for decoding. The backend is built based on the official REVERB corpus using the Kaldi script[3]. It is composed of a GMM-HMM system, a time-delay DNN (TDNN) trained with lattice-free maximum mutual information based on online-extracted i-vectors and MFCCs, and a tri-gram language model. Note that the window length and hop size for ASR are respectively 25 and 10 ms, following the default setup in Kaldi. During testing, we first obtain enhanced time-domain signals using our frontend and then feed them to the ASR backend for decoding, meaning that our frontend does not leverage any knowledge of the backend. We emphasize that the purpose of Test Set II and REVERB ASR is to show the generalization ability of our dereverberation models, which are trained based on simulated training data, as well as to compare the proposed algorithms with unsupervised methods such as WPE, not to obtain state-of-the-art performance using dereverberation frontends trained on the REVERB training data.

The two DNNs in Fig. 1 are trained sequentially. We first train the single-channel model using the first channel of all the multi-channel signals (in total 7,861 $\times$ 5 utterances). Designating the first microphone as the reference, we use the trained model to obtain a beamformed signal based on a random subset of microphones. The beamforming result is then combined with the mixture signal to train the second network. This way, the second DNN can deal with beamforming results produced by using up to eight microphones. Fig. 2 illustrates the DNN architecture. We use two-layer recurrent neural networks with bi-directional long short-term memory (BLSTM) having an encoder-decoder structure similar to U-Net, skip connections, and dense blocks as the learning machines for masking and mapping. The motivation for this DNN design is that BLSTM can model long-term temporal information, U-Net can maintain fine-grained local information as is suggested in image semantic segmentation [43], and DenseNet encourages feature reuse and improves the discriminative capability of the network [44]–[46]. In our experiments, this network architecture shows consistent improvements over the standard BLSTM and a recently

[2]DRR is computed as the energy ratio between the time-domain RIRs of direct-path signal and its reverberation.

[3][Online]. Available: https://github.com/kaldi-asr/kaldi/tree/master/egs/reverb/s5 (commit 61637e6c8ab01d3b4c54a50d9b20781a0aa12a59). Different from the Kaldi script, our study (1) performs sentence-level cepstral mean normalization on the input features of TDNN; (2) reduces the initial batch size of TDNN training by changing the trainer.num-chunk-per-minibatch flag from 256,128,64 to 128,64; (3) increases the number of TDNN training epochs from 10 to 20; (4) uses wsj/s5/local/wer_output_filter and wsj/s5/local/wer_hyp_filter to filter out tokens such as NOISE and SPOKEN_NOISE when utilizing local/score.sh to compute WER; and (5) enforces the same word insertion penalty (WIP) for near- and far-field conditions, and uses the averaged WER on the near- and far-field conditions of the validation set to select the best WIP.

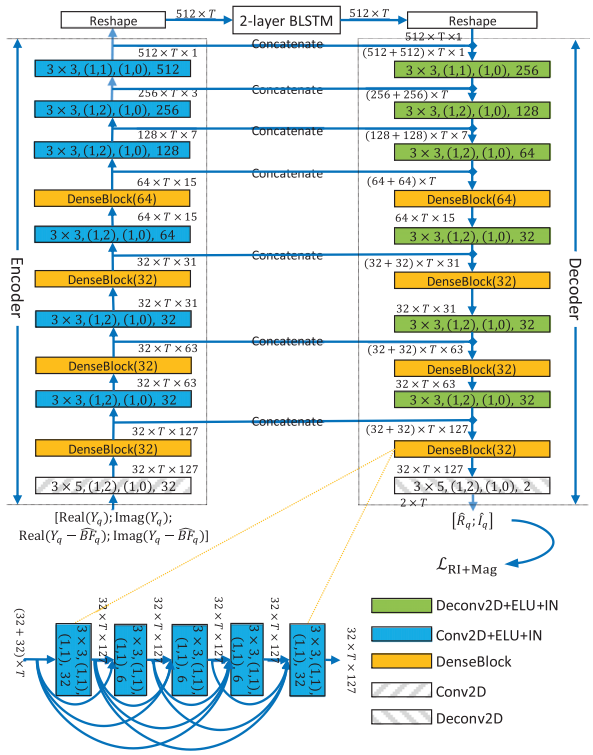

Fig. 2. Network architecture for predicting the RI components of S_q from the RI components of Y_q and $Y_q - \widehat{BF}_q$. Note that for single-channel processing, the network only takes in single-channel information as its inputs. The tensor shape after each block is in format: *featureMaps* × *timeSteps* × *frequencyChannels*. Each Conv2D, Deconv2D, Conv2D+ELU+IN, and Deconv2D+ELU+IN block is specified in format: *kernelSizeTime* × *kernelSizeFreq, (stridesTime, stridesFreq), (paddingsTime, paddingsFreq), featureMaps*. Each DenseBlock(g) contains five Conv2+ELU+IN blocks with growth rate g.

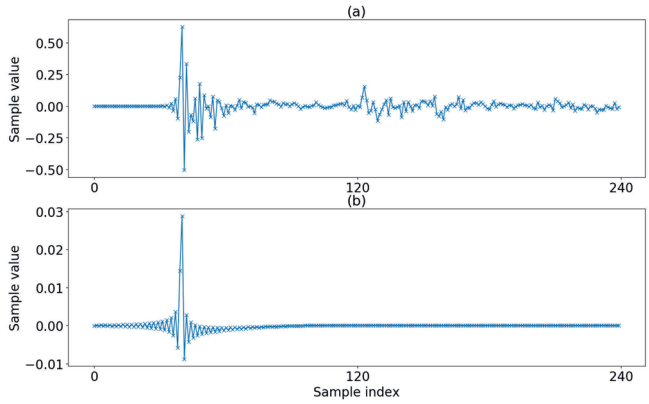

Fig. 3. RIR illustration. (a) Example RIR segment from REVERB (*RIR_SimRoom3_far_AnglB.wav*); (b) Example direct-path RIR simulated using RIR generator.

proposed convolutional recurrent neural network [22]. The encoder contains one two-dimensional (2D) convolution, and six convolutional blocks, each with 2D convolution, exponential linear units (ELUs) and instance normalization (IN) [47], for down-sampling. The decoder includes six convolutional blocks, each with 2D deconvolution, ELUs and IN, and one 2D deconvolution, for up-sampling. Each BLSTM layer has 512 units in each direction. The frontend processing uses 32 ms window length and 8 ms frame shift for STFT. The sampling rate is 16 kHz. A square-root Hann window is used as the analysis window.

Our main evaluation metrics are scale-invariant SDR (SI-SDR) [48] and PESQ, where the former is a time-domain metric that closely reflects the quality of estimated phase, and the latter strongly correlates with the accuracy of estimated magnitudes. We also consider scale-dependent SDR (SD-SDR) [48] for evaluating the single-channel models. Following REVERB, we also use cepstral distance (CD), log likelihood ratio (LLR), frequency-weighted segmental SNR (fwSegSNR), and speech-to-reverberation modulation energy ratio (SRMR) as the evaluation metrics. Note that the computation of SRMR does not require clean references. Word error rate (WER) is used to evaluate ASR performance.

B. Baseline Systems for Comparison

This section describes the single- and multi-channel baselines considered in our study.

- *Single-Channel Baselines*: The first four baselines for dereverberation perform single-channel magnitude-domain masking and mapping based magnitude spectrum approximation (MSA) and phase-sensitive spectrum approximation (PSA) [1], which are popular approaches in single-channel speech enhancement. We summarize the baselines in Table I. All of them use the same network architecture in Fig. 2, and the key difference is in the number of input and output feature maps depending on the input features and training targets, output non-linearities and loss functions. $T_a^b(\cdot) = \max(\min(\cdot, b), a)$ in $\mathcal{L}_{\mathrm{MSA-Masking}}$ and $\mathcal{L}_{\mathrm{PSA-Masking}}$ truncates the estimated mask to the range $[a, b]$. α in $\mathcal{L}_{\mathrm{MSA-Masking}}$ is set to 10.0, and β and γ in $\mathcal{L}_{\mathrm{PSA-Masking}}$ respectively set to 1.0 and 0.0 in our study.
- *TI-MVDR*: To show the effectiveness of using estimated complex spectra for covariance matrix computation, we apply the single-channel models to enhance each microphone signal following the last column of Table I, and then compute the covariance matrices based on Eq. (5) for TI-MVDR. This method is denoted as $\widehat{BF}_q$.
  Additionally, we use mask-weighted ways [28], [27] of computing covariance matrices for TI-MVDR, based on the estimated masks produced by the models trained with $\mathcal{L}_{\mathrm{MSA-Masking}}$ and $\mathcal{L}_{\mathrm{PSA-Masking}}$

$$\hat{\Phi}^{(d)}(f) = \frac{1}{T}\sum_{t} \eta^{(d)}(t,f)\, \boldsymbol{Y}(t,f)\, \boldsymbol{Y}(t,f)^{H}, \quad (10)$$

where $d \in \{s, v\}$.
When using $\mathcal{L}_{\mathrm{MSA-Masking}}$, $\eta^{(d)}$ is computed as

$$\eta^{(d)} = \mathrm{median}\left(\frac{T_0^{\alpha}\left(\hat{M}_1^{(d)}\right)}{T_0^{\alpha}\left(\hat{M}_1^{(s)}\right) + T_0^{\alpha}\left(\hat{M}_1^{(v)}\right)}, \ldots, \frac{T_0^{\alpha}\left(\hat{M}_P^{(d)}\right)}{T_0^{\alpha}\left(\hat{M}_P^{(s)}\right) + T_0^{\alpha}\left(\hat{M}_P^{(v)}\right)}\right), \quad (11)$$

TABLE I
SUMMARY OF VARIOUS SINGLE-CHANNEL MODELS FOR SPEECH DEREVERBERATION

Method	Input features	Loss function	Network Output	Output activation	Enhancement results
Complex spectral mapping	$\text{Real}(Y_q)$, $\text{Imag}(Y_q)$	$\mathcal{L}_{\text{RI}}$ or $\mathcal{L}_{\text{RI+Mag}}$	$\hat{R}_q, \hat{I}_q$	Linear	$\hat{S}_q = \hat{R}_q + j\hat{I}_q$ $\hat{V}_q = Y_q - \hat{S}_q$
MSA-Masking	$\lvert Y_q\rvert$	$\mathcal{L}_{\text{MSA-Masking}} = \left\| \lvert Y_q\rvert T_0^{\alpha}(\widehat{M}_q^{(s)}) - T_0^{\alpha\lvert Y_q\rvert}(\lvert S_q\rvert) \right\|_1 + \left\| \lvert Y_q\rvert T_0^{\alpha}(\widehat{M}_q^{(v)}) - T_0^{\alpha\lvert Y_q\rvert}(\lvert V_q\rvert) \right\|_1$	$\widehat{M}_q^{(s)}, \widehat{M}_q^{(v)}$	Clipped Softplus	$\hat{S}_q = Y_q T_0^{\alpha}(\widehat{M}_q^{(s)})$ $\hat{V}_q = Y_q T_0^{\alpha}(\widehat{M}_q^{(v)})$
MSA-Mapping		$\mathcal{L}_{\text{MSA-Mapping}} = \left\| \widehat{U}_q^{(s)} - \lvert S_q\rvert \right\|_1 + \left\| \widehat{U}_q^{(v)} - \lvert V_q\rvert \right\|_1$	$\widehat{U}_q^{(s)}, \widehat{U}_q^{(v)}$	Softplus	$\hat{S}_q = \widehat{U}_q^{(s)} e^{j\angle Y_q}$ $\hat{V}_q = \widehat{U}_q^{(v)} e^{j\angle Y_q}$
PSA-Masking		$\mathcal{L}_{\text{PSA-Masking}} = \left\| \lvert Y_q\rvert T_{\gamma}^{\beta}(\hat{Q}_q^{(s)}) - T_{\gamma\lvert Y_q\rvert}^{\beta\lvert Y_q\rvert}(\lvert S_q\rvert\cos(\angle S_q - \angle Y_q)) \right\|_1 + \left\| \lvert Y_q\rvert T_{\gamma}^{\beta}(\hat{Q}_q^{(v)}) - T_{\gamma\lvert Y_q\rvert}^{\beta\lvert Y_q\rvert}(\lvert V_q\rvert\cos(\angle V_q - \angle Y_q)) \right\|_1$	$\hat{Q}_q^{(s)}, \hat{Q}_q^{(v)}$	Sigmoid	$\hat{S}_q = Y_q T_{\gamma}^{\beta}(\hat{Q}_q^{(s)})$ $\hat{V}_q = Y_q T_{\gamma}^{\beta}(\hat{Q}_q^{(v)})$
PSA-Mapping		$\mathcal{L}_{\text{PSA-Mapping}} = \left\| \hat{Z}_q^{(s)} - \lvert S_q\rvert\cos(\angle S_q - \angle Y_q) \right\|_1 + \left\| \hat{Z}_q^{(v)} - \lvert V_q\rvert\cos(\angle V_q - \angle Y_q) \right\|_1$	$\hat{Z}_q^{(s)}, \hat{Z}_q^{(v)}$	Linear	$\hat{S}_q = \hat{Z}_q^{(s)} e^{j\angle Y_q}$ $\hat{V}_q = \hat{Z}_q^{(v)} e^{j\angle Y_q}$

where $\hat{M}_p^{(d)}$ denotes the estimated magnitude mask at microphone p.

When using $\mathcal{L}_{\text{PSA}-\text{Masking}}$, $\eta^{(d)}$ is computed as

$$\eta^{(d)} = \text{median}\left(T_{\gamma}^{\beta}\left(\hat{Q}_1^{(d)}\right), \ldots, T_{\gamma}^{\beta}\left(\hat{Q}_P^{(d)}\right)\right), \quad (12)$$

where $\hat{Q}_p^{(d)}$ denotes the estimated phase-sensitive mask at microphone p.

We also square the mask before median pooling, as the outer product is in the energy domain, while in Eq. (12) and (11) the mask is in the magnitude domain. $\eta^{(d)}$ is computed as

$$\eta^{(d)} = \text{median}\left(\frac{T_0^{\alpha}\left(\hat{M}_1^{(d)}\right)^2}{T_0^{\alpha}\left(\hat{M}_1^{(s)}\right)^2 + T_0^{\alpha}\left(\hat{M}_1^{(v)}\right)^2}, \ldots, \frac{T_0^{\alpha}\left(\hat{M}_P^{(d)}\right)^2}{T_0^{\alpha}\left(\hat{M}_P^{(s)}\right)^2 + T_0^{\alpha}\left(\hat{M}_P^{(v)}\right)^2}\right) \quad (13)$$

for $\mathcal{L}_{\text{PSA}-\text{Masking}}$ and as

$$\eta^{(d)} = \text{median}\left(T_{\gamma}^{\beta}\left(\hat{Q}_1^{(d)}\right)^2, \ldots, T_{\gamma}^{\beta}\left(\hat{Q}_P^{(d)}\right)^2\right) \quad (14)$$

for $\mathcal{L}_{\text{PSA}-\text{Masking}}$. Note that α, β and γ are respectively set to 10.0, 1.0 and 0.0 in our study.

- *Post-filtering (no re-training)*: After obtaining $\widehat{BF}_q$, we apply the single-channel models to $\widehat{BF}_q$ for post-filtering. The phase in $\widehat{BF}_q$ is used as the estimated phase for magnitude-domain masking and mapping based models. We emphasize that $\widehat{BF}_q$ is still very reverberant and is expected to contain low speech distortion. It is therefore reasonable to feed $\widehat{BF}_q$ into a single-channel model trained on unprocessed mixtures for further enhancement. In this method, only one DNN is trained (i.e., the single-channel model), but it is run twice at run time. This method is denoted as $\widehat{BF}_q$ + *Post-filtering (no re-training)*.
- *Post-filtering (re-training)*: As $\widehat{BF}_q$ may contain distortion unseen by the single-channel models, which are trained on unprocessed mixtures. We train a complex spectral mapping based post-filter, which predicts the RI components of S_q based on $\widehat{BF}_q$. Similar to the proposed system shown in Fig. 1, this method uses two DNNs, while the input to the second DNN is $\widehat{BF}_q$ rather than Y_q and $Y_q - \widehat{BF}_q$. We denote this method as $\widehat{BF}_q$ + *Post-filtering (re-training)*.
- *Single- and Multi-Channel WPE*: We follow the script for REVERB in the Kaldi toolkit, which is based on the open-source *nara-wpe* toolkit [49], to build our offline WPE baselines, where the window size is 32 ms and hop size is 8 ms, the prediction delay is set to 3, the iteration number set to 5, and the order of the regressive model set to 40 for single-channel processing and 10 for multi-channel processing. Note that these hyperparameters are the recommended ones in [16] and [6].

V. EVALUATION RESULTS

We first report the dereverberation performance of the trained models on Test Set I, and then report their generalization ability on Test Set II and REVERB ASR.

A. Dereverberation Performance on Test Set I

In Table II, we compare the performance of single-channel magnitude-domain masking and mapping based MSA and PSA, and complex spectral mapping over unprocessed speech and oracle magnitude-domain masks such as the spectral magnitude mask [50] and phase-sensitive mask [51]. Note that the unprocessed SI-SDR is closely related to DRR, an important factor characterizing the difficulty of dereverberation along with T60. Comparing $\mathcal{L}_{\text{MSA}-\text{Masking}}$, $\mathcal{L}_{\text{MSA}-\text{Mapping}}$, $\mathcal{L}_{\text{PSA}-\text{Masking}}$ and $\mathcal{L}_{\text{PSA}-\text{Mapping}}$ and $\mathcal{L}_{\text{RI}}$, we observe that $\mathcal{L}_{\text{RI}}$ leads to much better SI-SDR than MSA and PSA (6.2 vs. 0.8, 0.7, 2.3 and 1.6 dB), while MSA obtains the best PESQ (2.91 and 2.92 vs. 2.55, 2.56 and 2.80). This is likely because PESQ is closely related to the quality of estimated magnitudes, while time-domain measures such as SI-SDR needs the estimated magnitudes to compensate

TABLE II
AVERAGE SI-SDR (dB), PESQ AND SD-SDR (dB) OF DIFFERENT METHODS ON SINGLE-CHANNEL DEREVERBERATION (**TEST SET I**). ORACLE MASKING RESULTS ARE MARKED IN GRAY

Method	SI-SDR	PESQ	SD-SDR
Unprocessed	-3.7	1.93	-3.7
$\mathcal{L}_{\text{MSA-Masking}}$	0.8	2.91	3.5
$\mathcal{L}_{\text{MSA-Mapping}}$	0.7	2.92	3.5
$\mathcal{L}_{\text{PSA-Masking}}$	2.3	2.55	4.5
$\mathcal{L}_{\text{PSA-Mapping}}$	1.6	2.56	4.2
$\mathcal{L}_{\text{RI}}$	**6.2**	2.80	**7.2**
$\mathcal{L}_{\text{RI+Mag}}$	5.9	**3.07**	7.0
SMM ($T_0^{10}(\lvert S_q\rvert/\lvert Y_q\rvert)$)	1.6	3.40	3.9
PSM ($T_0^1(\lvert S_q\rvert\cos(\angle S_q - \angle Y_q)/\lvert Y_q\rvert)$)	4.5	3.09	5.8

TABLE III
AVERAGE SI-SDR (dB) AND PESQ OF DIFFERENT METHODS FOR TI-MVDR AND POST-FILTERING USING EIGHT MICROPHONES (**TEST SET I**)

Method	Model	Covariance Matrices	#mics	SI-SDR	PESQ
$\widehat{BF}_q$	$\mathcal{L}_{\text{MSA-Masking}}$	Eq. (5)	8	2.3	2.27
	$\mathcal{L}_{\text{MSA-Mapping}}$			2.3	2.26
	$\mathcal{L}_{\text{PSA-Masking}}$			3.3	2.31
	$\mathcal{L}_{\text{PSA-Mapping}}$			2.8	2.31
	$\mathcal{L}_{\text{RI}}$			5.8	2.34
	$\mathcal{L}_{\text{RI+Mag}}$			5.6	2.34
	$\mathcal{L}_{\text{MSA-Masking}}$	Eq. (10), (11)		1.7	2.44
	$\mathcal{L}_{\text{PSA-Masking}}$	Eq. (10), (12)		3.3	2.45
	$\mathcal{L}_{\text{MSA-Masking}}$	Eq. (10), (13)		3.0	2.44
	$\mathcal{L}_{\text{PSA-Masking}}$	Eq. (10), (14)		4.2	2.44
$\widehat{BF}_q$+*Post-filtering (no re-training)*	$\mathcal{L}_{\text{MSA-Masking}}$	Eq. (5)		4.4	3.01
	$\mathcal{L}_{\text{MSA-Mapping}}$			4.3	3.03
	$\mathcal{L}_{\text{PSA-Masking}}$			5.2	2.85
	$\mathcal{L}_{\text{PSA-Mapping}}$			4.7	2.87
	$\mathcal{L}_{\text{RI}}$			**9.6**	3.10
	$\mathcal{L}_{\text{RI+Mag}}$			9.4	**3.23**
	$\mathcal{L}_{\text{MSA-Masking}}$	Eq. (10), (11)		3.5	3.10
	$\mathcal{L}_{\text{PSA-Masking}}$	Eq. (10), (12)		5.3	2.96
	$\mathcal{L}_{\text{MSA-Masking}}$	Eq. (10), (13)		4.7	3.10
	$\mathcal{L}_{\text{PSA-Masking}}$	Eq. (10), (14)		6.1	2.95

for the error of phase estimation. Compared with $\mathcal{L}_{\text{RI}}$, $\mathcal{L}_{\text{RI+Mag}}$ substantially improves PESQ from 2.80 to 3.07, slightly degrading SI-SDR from 6.2 to 5.9 dB. In addition, $\mathcal{L}_{\text{RI+Mag}}$ obtains better PESQ than MSA (3.07 vs. 2.91 and 2.92), indicating the effectiveness of phase processing. We observe that SD-SDR results are consistent with SI-SDR. In the following experiments, we use $\mathcal{L}_{\text{RI+Mag}}$ as the loss function to train the two DNNs in Fig. 1, as it yields a very strong SI-SDR and the highest PESQ.

In Table III, we compare the performance of TI-MVDR and post-filtering based on the statistics computed using the single-channel models in Table II. Among all the alternative ways of computing the statistics for TI-MVDR, using the $\mathcal{L}_{\text{RI}}$ and $\mathcal{L}_{\text{RI+Mag}}$ models with Eq. (5) obtains the highest SI-SDR (5.8 and 5.6 dB), and the PESQ scores (2.34 and 2.34) are better than using MSA and PSA models with Eq. (5) (2.27, 2.26, 2.31 and 2.31) while worse than using MSA and PSA models with Eq. (10) (2.44, 2.45, 2.44 and 2.44). Applying post-filtering to $\widehat{BF}_q$ computed using the $\mathcal{L}_{\text{RI}}$ and $\mathcal{L}_{\text{RI+Mag}}$ models and Eq. (5) shows the highest SI-SDR scores (9.6 and 9.4 dB), and

TABLE IV
AVERAGE SI-SDR (dB) AND PESQ OF DIFFERENT METHODS ON MULTI-CHANNEL DEREVERBERATION (**TEST SET I**)

Metrics	#mics	Mixture	Model	$\widehat{BF}_q$+*Post-filtering (no re-training)*	$\widehat{BF}_q$+*Post-filtering (re-training)*	$\hat{S}_q^{(1)}$	$\hat{S}_q^{(2)}$
SI-SDR	1	-3.7	$\mathcal{L}_{\text{RI+Mag}}$ and Eq. (5)	-	-	5.9	-
	2			7.3	7.4	-	**7.5**
	3			8.2	8.9	-	**9.1**
	4			8.6	9.7	-	**9.9**
	6			9.2	10.6	-	**10.8**
	8			9.4	11.0	-	**11.2**
PESQ	1	1.93		-	-	3.07	-
	2			3.14	3.17	-	**3.18**
	3			3.20	**3.29**	-	**3.29**
	4			3.22	**3.34**	-	**3.34**
	6			3.23	3.40	-	**3.41**
	8			3.23	**3.44**	-	**3.44**

$\mathcal{L}_{\text{RI+Mag}}$ leads to significantly better PESQ over $\mathcal{L}_{\text{RI}}$ (3.23 vs. 3.10). These results suggest the effectiveness of complex spectral mapping based beamforming and post-filtering. In the following experiments, we compute $\widehat{BF}_q$ using Eq. (5) and $\mathcal{L}_{\text{RI+Mag}}$ if not specified, as this combination obtains the highest PESQ and a very competitive SI-SDR.

In Table IV, we show the results of $\hat{S}_q^{(2)}$, obtained by combining Y_q and $Y_q - \widehat{BF}_q$ for dereverberation (see Fig. 1). Consistently better performance is obtained over $\hat{S}_q^{(1)}$, confirming the effectiveness of multi-channel processing (e.g., 11.2 vs. 5.9 dB in SI-SDR and 3.44 vs. 3.07 in PESQ in the eight-microphone case). $\hat{S}_q^{(2)}$ also obtains better performance than $\widehat{BF}_q$ + *Post-filtering (no re-training)*, especially when the number of microphones is greater than two, for instance 11.2 vs. 9.4 dB in SI-SDR and 3.44 vs. 3.23 in PESQ in the eight-channel case. It is also slightly better than $\widehat{BF}_q$ + *Post-filtering (re-training)*. These results demonstrate the gains of combining $Y_q - \widehat{BF}_q$ with Y_q for dereverberation. In the two-channel case, it obtains results slightly better than $\widehat{BF}_q$ + *Post-filtering (no re-training)*, likely because $\widehat{BF}_q$ is not accurate enough in such a case. As a result, the quality of $Y_q - \widehat{BF}_q$ is not as good as when more microphones are available, and the trained DNN would focus on dealing with features computed from more than two microphones.

B. Generalization on Test Set II and REVERB ASR

In Table V, we directly evaluate the performance of the trained dereverberation models on Test Set II. Our models obtain dramatically better performance than WPE, and WPE + BeamformIt which applies weighted DAS (WDAS) beamforming on the output of WPE, and WPE + DNN-Based MVDR. Note that the first two baselines are available in Kaldi, and the third baseline applies DNN based TI-MVDR beamforming after WPE, where we use the single-channel model trained with $\mathcal{L}_{\text{RI+Mag}}$ and Eq. (5) to compute the statistics for MVDR, based on the signals processed after WPE. These comparisons show that the trained DNN models exhibit good generalization to novel reverberant and noisy conditions, and array configurations.

TABLE V
AVERAGE LLR, CD, FWSEGSNR, PESQ, AND SRMR OF DIFFERENT APPROACHES ON **TEST SET II**

Data	Metrics	Unprocessed	#mics	$\hat{S}_q^{(1)}$	$\hat{S}_q^{(2)}$	WPE	WPE+BeamformIt	WPE+DNN-Based MVDR ($\mathcal{L}_{\text{RI+Mag}}$ and Eq. (5))
SimData	CD	5.08	1	**3.16**	-	4.95	-	-
			2	-	**3.01**	4.98	4.66	4.77
			8	-	**2.78**	4.81	3.94	4.45
	LLR	0.67	1	**0.53**	-	0.63	-	-
			2	-	**0.45**	0.61	0.60	0.55
			8	-	**0.39**	0.53	0.49	0.40
	fwSegSNR	8.32	1	**15.61**	-	9.38	-	-
			2	-	**16.94**	9.71	10.20	11.24
			8	-	**18.75**	11.38	12.48	14.20
	PESQ	2.37	1	**3.29**	-	2.51	-	-
			2	-	**3.51**	2.58	2.65	2.77
			8	-	**3.71**	2.82	3.10	3.21
RealData	SRMR	3.18	1	**6.69**	-	3.83	-	-
			2	-	**6.38**	3.99	4.08	4.00
			8	-	**6.30**	5.04	5.53	5.29

TABLE VI
AVERAGE WER (%) OF DIFFERENT METHODS ON REAL DATA OF **REVERB ASR**

#mics	Method	Validation Set			Test Set		
		Near	Far	Avg	Near	Far	Avg
1	Mixture	16.53	17.22	16.88	17.31	17.05	17.18
	$\hat{S}_q^{(1)}$	**10.61**	**11.35**	**10.98**	**9.26**	**9.28**	**9.27**
	WPE	13.54	15.79	14.66	13.38	14.25	13.82
2	$\widehat{BF}_q$ ($\mathcal{L}_{\text{RI+Mag}}$ and Eq. (5))	21.21	22.83	22.02	21.02	18.26	19.64
	$\hat{S}_q^{(2)}$	**9.23**	**9.43**	**9.33**	**7.98**	**8.27**	**8.12**
	WPE	12.98	16.75	14.87	12.46	14.01	13.23
	WPE+BeamformIt	12.41	14.76	13.59	12.49	14.25	13.37
	WPE+DNN-Based MVDR ($\mathcal{L}_{\text{RI+Mag}}$ and Eq. (5))	16.91	20.98	18.95	17.18	14.01	15.59
8	$\widehat{BF}_q$ ($\mathcal{L}_{\text{RI+Mag}}$ and Eq. (5))	13.41	12.10	12.75	13.13	10.97	12.05
	$\hat{S}_q^{(2)}$	**7.92**	**7.72**	**7.82**	**5.88**	**6.41**	**6.14**
	WPE	12.48	15.31	13.89	11.21	11.75	11.48
	WPE+BeamformIt	9.54	10.59	10.06	8.24	8.61	8.43
	WPE+DNN-Based MVDR ($\mathcal{L}_{\text{RI+Mag}}$ and Eq. (5))	9.92	11.00	10.46	9.52	8.34	8.93

In Table VI, we report the ASR performance of the trained dereverberation models on the REVERB real data. The proposed approach obtains clear WER improvements over WPE, WPE + BeamformIt and WPE + DNN-Based MVDR (9.27% vs. 13.82% in the single-channel case, 8.12% vs. 13.23%, 13.37% and 15.59% in the two-channel case, and 6.14% vs. 11.48%, 8.43% and 8.93% in the eight-channel case). We observe large improvement by using $\hat{S}_q^{(2)}$, which can also be thought of as a variant of post-filtering, over $\widehat{BF}_q$. These results suggest that the trained dereverberation models can suppress reverberation with low speech distortion. We observe that the WPE + DNN-Based MVDR obtains better WER than $\widehat{BF}_q$, suggesting that WPE works as a frontend for DNN based beamforming, but worse WER than WPE + BeamformIt possibly because of the effects of reverberation.

VI. CONCLUSION

We have proposed a complex spectral mapping approach for speech dereverberation, where we predict the RI components of direct sound from the RI components of the mixture. We have extended this approach to address multi-channel dereverberation, by incorporating the RI components of cancelled speech for model training. Our single-channel and multi-channel dereverberation models show clear improvements over magnitude spectrum and phase-sensitive spectrum based models, and single- and multi-channel WPE. The trained models exhibit strong generalization to novel and representative reverberant environments and array configurations. Future research shall consider adaptive covariance matrix estimation, extensions to more noises, and online processing.

REFERENCES

[1] 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.
[2] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–774, Sep. 2009.
[3] S. Braun, B. Schwartz, S. Gannot, and E. A. P. Habets, "Late reverberation PSD estimation for single-channel dereverberation using relative convolutive transfer functions," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2016, pp. 1–5.
[4] 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.
[5] 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.
[6] M. Delcroix *et al.*, "Strategies for distant speech recognition in reverberant environments," *EURASIP J. Adv. Signal Process.*, vol. 60, 2015.
[7] E. A. P. Habets and P. A. Naylor, "Dereverberation," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. Hoboken, NJ, USA: Wiley, 2018, pp. 317–343.
[8] S. Braun *et al.*, "Evaluation and comparison of late reverberation power spectral density estimators," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 6, pp. 1056–1071, Jun. 2018.

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

[10] Z.-Q. Wang, K. Tan, and D. L. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 71–75.

[11] K. Han, Y. Wang, D. L. Wang, W. S. Woods, and I. Merks, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.

[12] M. Mimura, S. Sakai, and T. Kawahara, "Speech dereverberation using long short-term memory," in *Proc. Interspeech*, 2015, pp. 2435–2439.

[13] B. Wu *et al.*, "A reverberation-time-aware DNN approach leveraging spatial information for microphone array dereverberation," *EURASIP J. Adv. Signal Process.*, vol. 81, 2017.

[14] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *Proc. Eur. Signal Process. Conf.*, 2018, pp. 390–394.

[15] W. Mack, S. Chakrabarty, F.-R. Stöter, S. Braun, B. Edler, and E. A. P. Habets, "Single-channel dereverberation using direct MMSE optimization and bidirectional LSTM networks," in *Proc. Interspeech*, 2018, pp. 1314–1318.

[16] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation," in *Proc. Interspeech*, 2017, pp. 384–388.

[17] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Frame-online DNN-WPE dereverberation," in *Proceedings 16th Int. Workshop Acoust. Signal Enhancement*, 2018, pp. 466–470.

[18] J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, and T. Nakatani, "Joint optimization of neural network-based WPE dereverberation and acoustic model for robust online ASR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6655–6659.

[19] A. S. Subramanian, X. Wang, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita, "An investigation of end-to-end multichannel speech recognition for reverberant and mismatch conditions," 2019, *arXiv:1904.09049*.

[20] D. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[21] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2017, pp. 1–6.

[22] K. Tan and D. L. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, vol. 2019-May, pp. 6865–6869.

[23] 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.

[24] N. Roman, S. Srinivasan, and D. L. Wang, "Binaural segregation in multisource reverberant environments," *J. Acoust. Soc. Amer.*, vol. 120, pp. 4040–4051, 2006.

[25] U. Kjems and J. Jensen, "Maximum likelihood based noise covariance matrix estimation for multi-microphone speech enhancement," in *Proc. Eur. Signal Process. Conf.*, 2012, pp. 295–299.

[26] S. Wisdom *et al.*, "Differentiable consistency constraints for improved deep speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, vol. 2019, pp. 900–904.

[27] T. Yoshioka *et al.*, "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE Workshop Autom. Speech, Recognit. Understanding*, 2015, pp. 436–443.

[28] J. Heymann, L. Drude, A. Chinaev, and R. H.-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. IEEE Workshop Autom. Speech, Recognit. Understanding*, 2015, pp. 444–451.

[29] H. Erdogan *et al.*, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.

[30] X. Zhang, Z.-Q. Wang, and D. L. Wang, "A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust ASR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 276–280.

[31] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, "On time-frequency mask estimation for MVDR beamforming with application in robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 3246–3250.

[32] S. Gannot, E. Vincent, S. M.-Golan, and A. Ozerov, "A Consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[33] Y. Jiang, D. L. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 2112–2121, Dec. 2014.

[34] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time-frequency masks from spatial features," *Speech Commun.*, vol. 68, pp. 97–106, 2015.

[35] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 116–120.

[36] X. Zhang and D. L. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 5, pp. 1075–1084, May 2017.

[37] Z.-Q. Wang and D. L. Wang, "Combining spectral and spatial features for deep learning based blind speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 457–468, Feb. 2019.

[38] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5739–5743.

[39] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 1–5.

[40] K. Kinoshita *et al.*, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 7, no. 1, pp. 1–19, 2016.

[41] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE Challenge - Corpus description and performance evaluation," in *Proc. IEEE Workshop Appl. Signal Process., Audio, Acoust.*, 2015.

[42] E. A. P. Habets, "Room impulse response generator," 2010. [Online]. Available: https://github.com/ehabets/RIR-Generator/blob/master/rir_generator.pdf

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Compu. Assisted Intervention*, 2015, pp. 234–241.

[44] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 2261–2269.

[45] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. Int. Workshop Acoust. Signal Enhancement,* 2018, pp. 106–110.

[46] Y. Liu and D. L. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2092–2102, Dec. 2019.

[47] Y. Wu and K. He, "Group Normalization," in *proc. Eur. Conf. Comput. Vision,* 2018, pp. 3–19.

[48] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or well done?" in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 626–630.

[49] L. Drude, J. Heymann, C. Boeddeker, and R. H.-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *Proc. ITG Conf. Speech Commun.*, 2018, pp. 1–5.

[50] 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.

[51] 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.