Complex Spectral Mapping for Single- and Multi-Channel Speech Enhancement and Robust ASR

Zhong-Qiu Wang, *Student Member, IEEE*, Peidong Wang, *Graduate Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—This study proposes a complex spectral mapping approach for single- and multi-channel speech enhancement, where deep neural networks (DNNs) are used to predict the real and imaginary (RI) components of the direct-path signal from noisy and reverberant ones. The proposed system contains two DNNs. The first one performs single-channel complex spectral mapping. The estimated complex spectra are used to compute a minimum variance distortion-less response (MVDR) beamformer. The RI components of beamforming results, which encode spatial information, are then combined with the RI components of the mixture to train the second DNN for multi-channel complex spectral mapping. With estimated complex spectra, we also propose a novel method of time-varying beamforming. State-of-the-Art performance is obtained on the speech enhancement and recognition tasks of the CHiME-4 corpus. More specifically, our system obtains 6.82%, 3.19% and 1.99% word error rates (WER) respectively on the single-, two-, and six-microphone tasks of CHiME-4, significantly surpassing the current best results of 9.15%, 3.91% and 2.24% WER.**

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

I. Introduction

ENVIRONMENTAL noise and room reverberation are pervasive in modern hands-free speech communication applications such as digital assistants, teleconferencing and hearing aids. These kinds of acoustic interference are detrimental to modern automatic speech recognition (ASR) systems and dramatically degrade speech intelligibility and quality [1], [2]. Practical systems typically use multiple microphones to leverage spatial (in addition to spectral) information for speech enhancement and audio source separation. One common approach for multi-channel speech enhancement is beamforming followed by post-filtering [3], [4], where a popular method is to decompose a time-invariant or time-varying multi-channel Wiener filter into a product of a minimum variance distortion-less response (MVDR) beamformer and a real-valued post-filter. Conventionally, this approach requires an accurate estimate of target direction, and speech and noise power spectral density (PSD) and covariance matrices, which are typically computed based on sound localization such as GCC-PHAT [5], traditional speech enhancement [3], and blind source separation such as spatial clustering [6], [7]. Recently, DNN based time-frequency (T-F) masking or mapping have been established as a mainstream approach for speech enhancement and source separation [1]. Mask (or magnitude) estimation is dramatically improved using deep learning. Such real-valued mask estimates have been used to identify T-F units dominated by a single source, where the phase is less corrupted, for accurate source localization [8] and covariance matrix estimation [9], [10]. All the top teams in the recent CHiME-4 challenge adopted T-F masking and deep learning based beamforming in their ASR systems [10].

We investigate single- and multi-channel DNN-based speech enhancement and robust ASR. In addition to mask (or magnitude) estimation, our study explores the effects of phase estimation for multi-channel speech enhancement. We emphasize that current T-F masking based approaches for beamforming typically compute spatial covariance matrices as a summation of mixture outer products weighted by a mask [6], [9], [11]–[14]. In environments with strong noise and room reverberation, there may be insufficient T-F units dominated by target speech, and the mixture outer product at each T-F unit inevitably contains noise and reverberation. We believe, in such cases, that it is beneficial to perform phase estimation in addition to magnitude estimation and directly use the estimated complex spectra for covariance matrix computation. This method is simpler as it does not involve a weighting mechanism. In addition, real-valued post-filtering only performs magnitude estimation and would inevitably produce phase inconsistency issues [15]–[17], i.e. no time-domain signal corresponds to the estimated complex spectrogram. Although beamforming typically improves phase, its performance heavily depends on the number of microphones and is susceptible to strong room reverberation [3]. Phase estimation would hence be needed for post-filtering in order to further improve the phase produced by beamforming. Although modern ASR systems only consider magnitude-based features such as log Mel features, accurate phase estimation can indirectly benefit ASR as better estimated phase leads to better spatial processing such as beamforming and target localization.

Manuscript received January 9, 2020; revised April 18, 2020; accepted May 21, 2020. Date of publication May 28, 2020; date of current version June 18, 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 the Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Yu Tsao. *(Z.-Q. Wang and P. Wang made equal contributions to this paper.) (Corresponding author: Zhong-Qiu Wang.)*

Zhong-Qiu Wang and Peidong Wang are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: wangzhon@cse.ohio-state.edu; wang.7642@osu.edu).

DeLiang Wang is with the Department of Computer Science and Engineering & 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.2020.2998279

Our study performs DNN based phase estimation and investigates its effects on single-channel enhancement, time-invariant and time-varying beamforming, and post-filtering. We perform speech enhancement in the complex domain [18], more specifically via complex spectral mapping [19], [20], which was originally proposed to deal with single-channel speech enhancement in anechoic conditions. This paper goes beyond previous work on complex spectral mapping by using a new loss function and addressing multi-channel speech enhancement and robust ASR. Our contributions can be summarized as follows. First, we introduce a magnitude-domain loss for complex spectral mapping, which leads to better enhancement results especially in terms of PESQ. Second, complex spectra produced by single- and multi-channel complex spectral mapping are found to produce slightly better signal statistics for beamforming than magnitude-domain mask estimates, especially in relatively matched conditions. Third, a novel way of using estimated complex spectra for time-varying beamforming is proposed. Fourth, the proposed system advances state-of-the-art enhancement and recognition results on the single-, two- and six-microphone tasks of CHiME-4, without using any model ensemble as employed in the previous best results reported in [21] and [22] that combines multiple frontends and backends. It should be noted that preliminary versions of this study [13], [23]–[26] have been presented in ICASSP 2017 and 2018, but this paper employs complex spectral mapping for the first time and obtains much better results.

The rest of this paper is organized as follows. We describe our physical model and objectives in Section II, and present the proposed algorithms in Section III. Experimental setup and evaluation results are presented in Section IV and V. Conclusions are made in Section VI.

II. Physical Model and Objectives

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

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

where $S_q(t,f) \in \mathbb{C}$ is the complex STFT coefficient of the direct-path signal captured by a reference microphone q at time t and frequency f, $\boldsymbol{c}(f;q) \in \mathbb{C}^{P\times 1}$ denotes 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, its reverberation, reverberant noise and captured mixture. The target speaker is assumed still (non-moving) within each utterance.

Our study proposes multiple deep learning based algorithms to extract the direct-path signal S_q from the mixture Y_q captured at reference microphone q, with or without exploiting spatial information contained in $\boldsymbol{Y}$. We assume offline processing scenarios. We normalize the sample variance of each time-domain mixture to one before any processing. This normalization can deal with random gains in input signals, and hence would be important for mapping-based methods for speech enhancement. In addition, our network architecture (see Fig. 2) has short-cut connections from network input to output. They can help mapping based methods determine the gain of target speech.

The proposed algorithms are designed such that the models, once trained, can be readily applied to arrays with any number of microphones arranged in an unknown geometry. This flexibility is useful for cloud-based services, where client devices vary in the number of microphones and microphone geometry, but poses challenges for supervised learning based approaches, as they require fixed input and output dimensions and may have limited generalization capability to a new array geometry.

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

III. Proposed Algorithms

Fig. 1 shows two DNNs in the proposed system. The first one performs single-channel complex spectral mapping based enhancement. The enhancement results are utilized to compute signal statistics for an MVDR beamformer. The beamforming results are combined with the mixture for the second DNN to perform multi-channel complex spectral mapping based post-filtering so that spectral and spatial information can be integrated for DNN training. A second beamformer is then computed for speech recognition, as the second DNN can produce better signal statistics for beamforming after leveraging spatial information. In Fig. 1, the multiple MVDR beamformers after the first DNN differ in the choice of reference microphone. This is because the direct-path signals at different microphones vary a lot in terms of phase, and we need to set each one of the microphones as the reference in MVDR beamforming to estimate each direct-path signal.

A. Single-Channel Complex Spectral Mapping

Following [18]–[20], we train a DNN to directly predict the RI components of the direct-path signal from noisy and reverberant ones. We use the following loss function for model training

$$\mathcal{L}_{\text{RI}} = \left\|\hat{R}_p - \text{Real}\left(S_p\right)\right\|_1 + \left\|\hat{I}_p - \text{Imag}\left(S_p\right)\right\|_1 \tag{2}$$

where $p \in \{1,\ldots,P\}$ indexes microphones, $\hat{R}_p$ and $\hat{I}_p$ are the predicted RI components, and $\text{Real}(\cdot)$ and $\text{Imag}(\cdot)$ extract the RI components. The enhancement result at microphone p is computed as $\hat{S}_p^{(k)} = \hat{R}_p^{(k)} + j\hat{I}_p^{(k)}$, where j is the imaginary unit. As the overall system has two DNNs (see Fig. 1), we use superscript $k \in \{1,2\}$ to denote that the output is produced by the k^{th} DNN.

Following recent studies [19], [27] that include a magnitude-domain loss for complex spectra approximation, we design the following loss function

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

The motivation is that using $\mathcal{L}_{\text{RI}}$ alone does not lead to satisfactory magnitude estimates, as the estimated magnitudes need to compensate for the estimation error made on phase. A major difference from [19], [27] is that we do not perform power or logarithmic compression on the magnitude spectra for loss

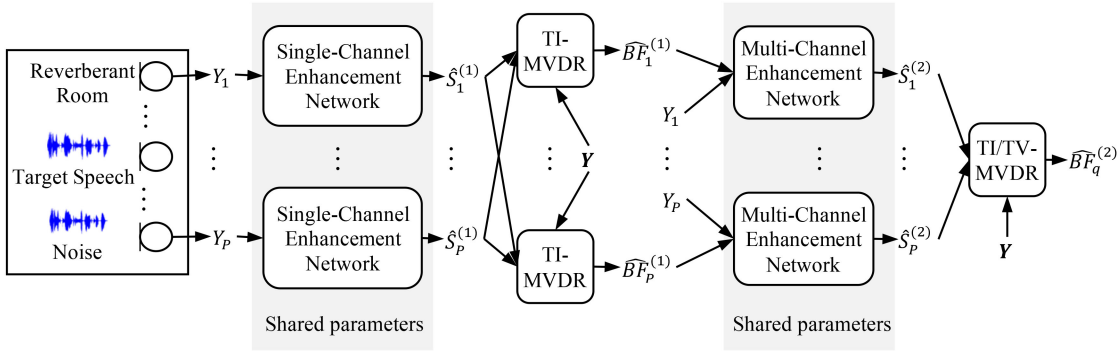


Fig. 1. System diagram of overall system for single- and multi-channel speech enhancement. There are two DNNs, one taking in single-channel and the other multi-channel information for speech enhancement. The superscripts in $\hat{S}_p^{(1)}$ and $\widehat{BF}_p^{(1)}$, and $\hat{S}_p^{(2)}$ and $\widehat{BF}_p^{(2)}$ for $p \in \{1, \ldots, P\}$ respectively denote whether they are produced by the first and the second DNN. The MVDR beamformer can be time-invariant (TI-MVDR) or time-varying (TV-MVDR). Detailed DNN architecture is shown in Fig. 2.

computation. This way, the DNN is always trained to estimate a spectrogram that has consistent phase and magnitude structure, and hence would likely produce a consistent STFT spectrogram at run time [17].

B. Multi-Channel Complex Spectral Mapping

After obtaining $\hat{S}_p^{(1)}$ at each microphone using single-channel complex mapping, we directly use the estimated complex spectra, rather than estimated T-F masks as a weighting mechanism [6], [9], [11]–[14], for covariance matrix computation

$$\hat{\Phi}^{(s)}(f) = \frac{1}{T}\sum_{t=1}^{T} \hat{\boldsymbol{S}}(t,f)\, \hat{\boldsymbol{S}}(t,f)^H \tag{4}$$

$$\hat{\Phi}^{(v)}(f) = \frac{1}{T}\sum_{t=1}^{T} \hat{\boldsymbol{V}}(t,f)\, \hat{\boldsymbol{V}}(t,f)^H \tag{5}$$

where $\hat{\boldsymbol{V}} = \boldsymbol{Y} - \hat{\boldsymbol{S}}$ and T is the total number of frames in the utterance. The rationale is that the estimated complex spectra produced by complex spectral mapping are expected to have better phase than the mixture, and hence could lead to better estimation of covariance matrices.

The relative transfer function is then computed as

$$\begin{aligned} \hat{\boldsymbol{r}}(f) &= \mathcal{P}\left\{\hat{\Phi}^{(s)}(f)\right\} \\ \hat{\boldsymbol{c}}(f;q) &= \hat{\boldsymbol{r}}(f)/\hat{r}_q(f) \end{aligned} \tag{6}$$

where $\mathcal{P}\{\cdot\}$ extracts the principal eigenvector and $\hat{r}_q(f)$ denotes the q^{th} element of $\hat{\boldsymbol{r}}(f)$. Note that the speech covariance matrix, if accurately estimated, is close to a rank-one matrix for still directional sources. Its principal eigenvector is hence a reasonable estimate of the steering vector [11], [13], [3]. $\hat{\boldsymbol{r}}(f)$ is then divided by $\hat{r}_q(f)$ to obtain the relative transfer function with respect to microphone q. Without this operation, MVDR beamforming would introduce a random complex gain at each frequency, leading to speech distortion in the beamformed mixture.

A time-invariant MVDR (TI-MVDR) is then computed using

$$\hat{\boldsymbol{w}}(f;q) = \frac{\hat{\Phi}^{(v)}(f)^{-1}\hat{\boldsymbol{c}}(f;q)}{\hat{\boldsymbol{c}}(f;q)^H \hat{\Phi}^{(v)}(f)^{-1}\hat{\boldsymbol{c}}(f;q)} \tag{7}$$

The beamformed mixture is obtained as $\widehat{BF}_q(t,f) = \hat{\boldsymbol{w}}(f;q)^H \boldsymbol{Y}(t,f)$.

The second DNN takes in the RI components of $\widehat{BF}_q$ and Y_q to predict the RI components of S_q. Here $\widehat{BF}_q$ can be considered as a spatial feature [28]–[32], to guide the DNN to extract a target speech signal in a particular direction and with specific spectral structure. In addition, it is in the complex domain, and therefore could help improve phase estimation in addition to enhance magnitude estimation. In contrast, previously proposed spatial features, such as inter-channel phase differences (IPDs) [33], cosine and sine IPDs [29], [34], target direction compensated IPDs [25], and the magnitudes of beamformed mixtures [32], are in the real domain and only used for improving magnitude estimation.

Here we consider MVDR beamforming, as it is expected to introduce very little speech distortion. Such beamforming results could help the second DNN better predict the direct-path signal. Other beamformers, such as multi-channel Wiener filter (MCWF) and generalized eigenvector (GEV) beamforming, essentially use the same projection direction (i.e. $\hat{\Phi}^{(v)}(f)^{-1}\hat{\boldsymbol{c}}(f;q)$) as MVDR, but use different spectral gains for further noise suppression [3]. Such gains, which are real-valued for MCWF and complex-valued for GEV, introduce speech distortion. In addition, such gains are typically computed based on signal statistics, but better gains can likely be produced using DNNs. As a result, our study considers distortionless MVDR beamforming, and uses DNN-based post-filtering for further noise suppression.

C. Adaptive Covariance Matrix Computation

Since the target speaker is typically still within each utterance, it is reasonable to estimate RTF from $\hat{\Phi}^{(s)}(f)$ using all the frames within an utterance. Clearly, more frames in this case lead to more accurate RTF estimation for a still directional source. However, even if the target speaker is still, the spatial coherence of environmental noise and room reverberation can be highly time-varying in real-world environments such as the BUS and CAF conditions in the CHiME-4 corpus. It is hence necessary to estimate noise covariance matrix per T-F unit or per block of units rather than per frequency for better noise suppression.

We follow a recently proposed algorithm [35] to estimate time-varying noise covariance matrices. In [35], per-frequency T-F mask based covariance matrix is considered as a prior, and under a maximum a posterior framework, the time-varying spatial covariance matrix at each T-F unit is computed as a weighted combination of the prior and the summation of the mask-weighted mixture outer products in each non-overlapping block of T-F units. Specifically, we compute the time-varying noise covariance matrix in the following way

$$\hat{\Phi}^{(v)}(t,f) = (1-\alpha)\frac{\sum_{t-\Delta}^{t+\Delta}\hat{\boldsymbol{V}}(t,f)\hat{\boldsymbol{V}}(t,f)^H}{trace\left(\sum_{t-\Delta}^{t+\Delta}\hat{\boldsymbol{V}}(t,f)\hat{\boldsymbol{V}}(t,f)^H\right)\Big/ P} + \alpha\frac{\hat{\Phi}^{(v)}(f)}{trace\left(\hat{\Phi}^{(v)}(f)\right)\Big/ P} \tag{8}$$

where α is empirically set to 0.5 and Δ is half the window size in frames. Different from [35], we use estimated complex spectra produced by complex spectral mapping, rather than estimated masks in a mask-weighted fashion, for covariance matrix computation. This could result in more accurate covariance estimation. In addition, we normalize the energy levels before the weighted sum to eliminate the effects of time-varying PSD and focus on the weighted summation of spatial coherences, as noise PSD cancels out in MVDR beamforming. Without the energy normalization, the summation can be easily dominated by one of the two terms, since noise PSD can be highly non-stationary. We emphasize that the first term is computed based on a small context window of $2\Delta+1$ frames, while the second term based on all the frames. This way, the computation of the noise covariance matrix can leverage long-term stationary information and, at the same time, adapt to sudden changes of noise characteristics. Note that the short-term noise covariance matrix needs an accurate complex spectrum estimate, which is obtained using complex spectral mapping.

A time-varying MVDR (TV-MVDR) beamformer is then computed as

$$\hat{\boldsymbol{w}}(t,f;q) = \frac{\hat{\Phi}^{(v)}(t,f)^{-1}\hat{\boldsymbol{c}}(f;q)}{\hat{\boldsymbol{c}}(f;q)^H\hat{\Phi}^{(v)}(t,f)^{-1}\hat{\boldsymbol{c}}(f;q)} \tag{9}$$

and the beamforming result is computed using $\widehat{BF}_q(t,f) = \hat{\boldsymbol{w}}(t,f;q)^H\boldsymbol{Y}(t,f)$.

After cross validation, Δ is set to 0 for the two-microphone and 3 for the six-microphone recognition task of CHiME-4. We found that setting Δ to 0 always produces more noise suppression and the beamformed signals always sound better, simply because we use fine-grained T-F unit level noise estimates produced by DNN to compute $\hat{\Phi}^{(v)}(t,f)$. However, such noise estimates may not be perfectly accurate. In the six-channel case, MVDR has more degrees of freedom to minimize the energy of such unperfect noise estimates, but at a risk of introducing more speech distortion. We therefore slightly increase Δ to have more stable $\hat{\Phi}^{(v)}(t,f)$ and sacrifice noise reduction for the reduction of speech distortion. In the two-microphone case, even if we set Δ to 0, the noise reduction is limited (it only has one degree of freedom for nosie reduciton) and hence speech distortion is little.

IV. Experimental Setup

We evaluate our algorithms on the enhancement and recognition tasks of the publicly-available CHiME-4 corpus [10], a popular dataset featuring one-, two- and six-microphone tasks designed for robust ASR. Our study always considers the direct outputs from DNN (i.e. $\hat{S}_q^{(1)}$ and $\hat{S}_q^{(2)}$) for speech enhancement, and beamforming results (i.e. $\widehat{BF}_q^{(1)}$ and $\widehat{BF}_q^{(2)}$) for speech recognition, as it is well-known that beamforming produces less speech distortion, which is important for modern ASR systems, but also less noise reduction, compared to deep learning based masking and mapping. This section details the CHiME-4 dataset, our proposed frontend and several baseline frontends, and our ASR backend.

A. CHiME-4 Corpus

The CHiME-4 corpus [10] contains six-microphone simulated and real recordings. The microphones are mounted on a tablet, with five of them facing the front and the other one facing the rear. This corpus contains recordings from four real-world environments (including street, pedestrian areas, cafeteria and bus), exhibiting large training and testing mismatches in terms of speaker, noise and spatial characteristics, and around 12% of the real recordings suffer from microphone failures. The training data includes 7,138 simulated and 1,600 recorded utterances, the validation data contains 1,640 simulated and 1,640 recorded utterances, and the test data consists of 1,320 simulated and 1,320 recorded utterances. Each of the three recorded datasets is constructed using four different speakers. It should be noted that reverberation is weak in the CHiME-4 corpus, partly because the considered environments are not very reverberant and the speaker-microphone distance is not large for a hand-held position. The single-channel task uses one of the six microphones for testing. For the two-microphone task, two of the front five channels that do not suffer from microphone failure are selected for each utterance for testing. To address microphone failures in the real recordings of the six-microphone task, we first select a microphone signal that is most correlated with the other five, and then throw away the signals with less than 0.3 correlation coefficients with the selected signal.

B. Frontend Enhancement System

We use all the simulated signals in the training set to train our frontends, and report the enhancement results on the simulated test set. We consider the clean signal captured by the fifth microphone as the reference for metric computation, since it exhibits the highest signal-to-noise ratio among all the microphones.

The two DNNs in Fig. 1 are trained sequentially. After training the first DNN, we use it to generate for each microphone a beamformed signal based on TI-MVDR and a random number of microphones, leading to 7,138$*$6 beamformed signals in total. Each beamformed signal is combined with the mixture signal to train the second DNN. This way, the second DNN

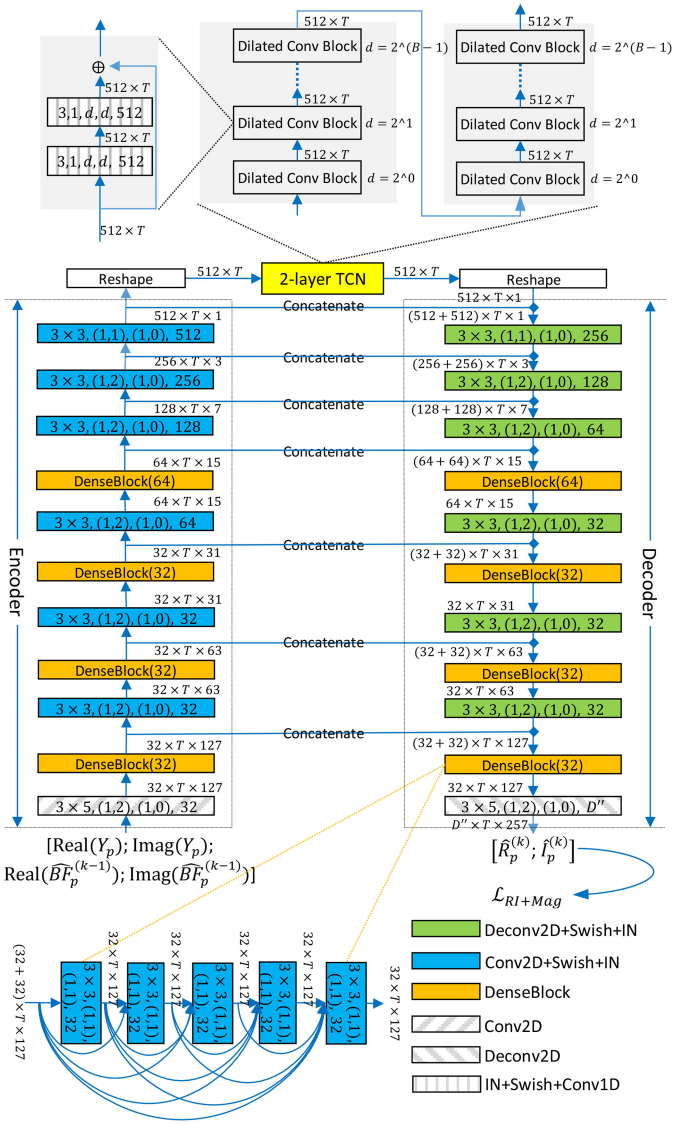


Fig. 2. Network architecture for predicting the RI components of S_q from the RI components of Y_q and $\widehat{BF}_q$. For single-channel processing, the network only takes single-channel information as its inputs. The tensor shape after each encoder-decoder block is in the format: *featureMaps* × *timeSteps* × *frequencyChannels*. Each of Conv2D, Deconv2D, Conv2D + IN + Swish, and Deconv2D + IN + Swish blocks is specified in the format: *kernelSizeTime* × *kernelSizeFreq, (stridesTime, stridesFreq), (paddingsTime, paddingsFreq), featureMaps*. Each DenseBlock(g) contains five Conv2D + IN + Swish blocks with growth rate g. The tensor shape after each TCN block is in the format: *featureMaps* × *timeSteps*. Each IN + Swish + Conv1D block is specified in the format: *kernelSizeTime, stridesTime, paddingsTime, dilationTime, featureMaps*.

can deal with the TI-MVDR results produced by using up to six microphones. In our experiments, we also tried using TV-MVDR to produce beamformed signals for training the second DNN. The performance is however not clearly better. This could be because TI-MVDR uses the same filter per frequency along time and can therefore produce stable spectral patterns conducive to training convolutional networks.

The network architecture for enhancement is shown in Fig. 2. The network is a temporal convolutional network (TCN) [36] with encoder-decoder structure similar to U-Net [37], skip connections, and dense blocks [38], [39]. The motivation for this network design is that TCN can model long-term temporal dependencies through large receptive fields achieved via dilated convolution, U-Net can maintain fine-grained local spectral structure as suggested in image semantic segmentation [37], and dense blocks can increase feature reuse and improve the discriminative power of the network [38]. A similar architecture was recently used in a state-of-the-art speaker separation algorithm [40]. The encoder contains one two-dimensional (2D) convolution, and six convolutional blocks, each with 2D convolution, Swish non-linearity and instance normalization (IN), for down-sampling. The decoder includes six blocks of 2D deconvolution, Swish and IN, and one 2D deconvolution, for up-sampling. The TCN contains two layers, each of which has six dilated convolutional blocks. We use two one-dimensional (1D) depth-wise separable convolution in each dilated convolutional block to reduce the number of parameters. Each model has around 13 million parameters.

The frame length is 32 ms and frame shift 8 ms. Square-root Hann window is used as the analysis window. The sampling rate is 16 kHz. A 512-point discrete Fourier transform is used to extract complex STFT spectrograms. No global mean-variance normalization is performed on the input features. For complex spectral mapping, linear activation is used in the output layer to produce estimated RI components. As the CHiME-4 dataset exhibits diverse gains at different microphones, we separately normalize each of the six microphone signals to have unit sample variance before any frontend processing.

We use PESQ, STOI, scale-invariant signal-to-distortion ratio (SI-SDR) [41], and BSS-Eval SDR as the evaluation metrics. PESQ and STOI strongly correlate with the accuracy of estimated magnitude. On the other hand, SI-SDR is a time-domain metric closely reflecting the quality of estimated magnitude and phase, meaning that magnitude estimates need to compensate for the inaccuracy of phase estimates in order to produce a high SI-SDR.

C. Baseline Frontend Systems

We consider four single-channel benchmarks listed in Table I to demonstrate the effectiveness of single-channel complex spectral mapping based speech enhancement. The four benchmarks are based on masking and mapping based magnitude spectrum approximation (MSA) [1] and phase-sensitive spectrum approximation (PSA) [42]. All of them use the same network architecture as shown in Fig. 2. The main differences lie in the number of input and output feature maps, and the activation function in the output layer. In $\mathcal{L}_{\text{MSA}-\text{Masking}}$ and $\mathcal{L}_{\text{PSA}-\text{Masking}}$, $T_a^b(\cdot) = \max(\min(\cdot, b), a)$ truncates the estimated masks to the range $[a, b]$. β is set to 5.0 in $\mathcal{L}_{\text{MSA}-\text{Masking}}$ and γ set to 1.0 in $\mathcal{L}_{\text{PSA}-\text{Masking}}$.

In addition, we investigate the effectiveness of the single-channel models for TI-MVDR beamforming. One way is to apply each single-channel model to each microphone signal to obtain $\hat{\boldsymbol{S}}$ and $\hat{\boldsymbol{V}}$, perform TI-MVDR beamforming using Eq. (4)–(7), and compare their ASR performance. This comparison

TABLE I
SUMMARY OF SINGLE-CHANNEL FRONTENDS

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	$\lvert Y_q\rvert$	$\mathcal{L}_{\text{MSA-Masking}} = \left\lVert \lvert Y_q\rvert T_0^{\beta}(\widehat{M}_q^{(s)}) - T_0^{\beta\lvert Y_q\rvert}(\lvert S_q\rvert)\right\rVert_1 + \left\lVert \lvert Y_q\rvert T_0^{\beta}(\widehat{M}_q^{(v)}) - T_0^{\beta\lvert Y_q\rvert}(\lvert V_q\rvert)\right\rVert_1$	$\widehat{M}_q^{(s)}, \widehat{M}_q^{(v)}$	Clipped Softplus	$\hat{S}_q = Y_q T_0^{\beta}(\widehat{M}_q^{(s)})$ $\hat{V}_q = Y_q T_0^{\beta}(\widehat{M}_q^{(v)})$
MSA-Mapping		$\mathcal{L}_{\text{MSA-Mapping}} = \left\lVert \hat{R}_q^{(s)} - \lvert S_q\rvert\right\rVert_1 + \left\lVert \hat{R}_q^{(v)} - \lvert V_q\rvert\right\rVert_1$	$\hat{R}_q^{(s)}, \hat{R}_q^{(v)}$	Softplus	$\hat{S}_q = \hat{R}_q^{(s)} e^{j\angle Y_q}$ $\hat{V}_q = \hat{R}_q^{(v)} e^{j\angle Y_q}$
PSA-Masking		$\mathcal{L}_{\text{PSA-Masking}} = \left\lVert \lvert Y_q\rvert T_0^{\gamma}(\hat{Q}_q^{(s)}) - T_0^{\gamma\lvert Y_q\rvert}(\lvert S_q\rvert\cos(\angle S_q - \angle Y_q))\right\rVert_1 + \left\lVert \lvert Y_q\rvert T_0^{\gamma}(\hat{Q}_q^{(v)}) - T_0^{\gamma\lvert Y_q\rvert}(\lvert V_q\rvert\cos(\angle V_q - \angle Y_q))\right\rVert_1$	$\hat{Q}_q^{(s)}, \hat{Q}_q^{(v)}$	Sigmoid	$\hat{S}_q = Y_q T_0^{\gamma}(\hat{Q}_q^{(s)})$ $\hat{V}_q = Y_q T_0^{\gamma}(\hat{Q}_q^{(v)})$
PSA-Mapping		$\mathcal{L}_{\text{PSA-Mapping}} = \left\lVert \hat{Z}_q^{(s)} - \lvert S_q\rvert\cos(\angle S_q - \angle Y_q)\right\rVert_1 + \left\lVert \hat{Z}_q^{(v)} - \lvert V_q\rvert\cos(\angle V_q - \angle Y_q)\right\rVert_1$	$\hat{Z}_q^{(s)}, \hat{Z}_q^{(v)}$	Linear	$\hat{S}_q = \hat{Z}_q^{(s)} e^{j\angle Y_q}$ $\hat{V}_q = \hat{Z}_q^{(v)} e^{j\angle Y_q}$

can show the effectiveness of single-channel phase estimation when its result is used for beamforming.

We also evaluate the mask weighting technique for collecting statistics for TI-MVDR beamforming, based on the MSA-Masking and PSA-Masking models. Following [6], [9], [12], [13], we compute the covariance matrices in the following way

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

where $d \in \{s, v\}$, and $\eta^{(d)}$ is computed as

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

for MSA-Masking and as

$$\eta^{(d)} = \text{median}\left(T_0^{\gamma}\left(\hat{Q}_1^{(d)}\right), \ldots, T_0^{\gamma}\left(\hat{Q}_P^{(d)}\right)\right) \tag{12}$$

for PSA-Masking. β in Eq. (11) and γ in Eq. (12) are respectively set to 5.0 and 1.0. We emphasize that the six-channel task of CHiME-4 contains recordings with microphone failure. According to [9], the median pooling operation is an effective way of addressing microphone failure. Thus, we also compute covariance matrices based on the estimated spectra produced by complex spectral mapping in the following way

$$\hat{\Phi}^{(s)}(f) = \frac{1}{T}\sum_{t=1}^{T} \eta^{(s)}(t,f)\hat{\boldsymbol{S}}(t,f)\ \hat{\boldsymbol{S}}(t,f)^H \tag{13}$$

$$\hat{\Phi}^{(v)}(f) = \frac{1}{T}\sum_{t=1}^{T} \eta^{(v)}(t,f)\hat{\boldsymbol{V}}(t,f)\, \hat{\boldsymbol{V}}(t,f)^H \tag{14}$$

where $\eta^{(s)}$ and $\eta^{(v)}$ are computed as

$$\eta^{(s)} = \text{median}\left(\frac{\left|\hat{S}_1\right|}{\left|\hat{S}_1\right| + \left|\hat{V}_1\right|}, \ldots, \frac{\left|\hat{S}_P\right|}{\left|\hat{S}_P\right| + \left|\hat{V}_P\right|}\right) \tag{15}$$

$$\eta^{(v)} = \text{median}\left(\frac{\left|\hat{V}_1\right|}{\left|\hat{S}_1\right| + \left|\hat{V}_1\right|}, \ldots, \frac{\left|\hat{V}_P\right|}{\left|\hat{S}_P\right| + \left|\hat{V}_P\right|}\right) \tag{16}$$

Different from Eq. (10), Eqs. (13) and (14) use estimated speech and noise complex spectra for covariance matrix computation, while Eq. (10) uses mixture complex spectra. Note that we only apply this technique to the six-microphone recognition task, as the two-microphone recognition task and the enhancement task do not contain signals with microphone failure.

D. Backend Recognition System

Our ASR backend is a DNN-HMM hybrid system. The acoustic model is trained using both simulated and recorded noisy utterances in the training set. The input features to the acoustic model are 80-dimensional logarithmically compressed Mel filterbank feature together with its delta and double delta. The acoustic model is a wide-residual BLSTM network (WRBN) [43] trained with utterance-wise recurrent dropout [26]. At test time, we perform lattice re-scoring using the task-standard trigram, five-gram and RNN language models, and an LSTM language model (LSTMLM) recently proposed in [44]. The LSTMLM re-scored lattice is used for unsupervised speaker adaptation. We apply iterative speaker adaptation proposed in [26] for three iterations, each of which uses the LMSTLM re-scored lattice to fine-tune a linear input layer [45] prepended to the acoustic model.

Since the ASR system uses different frame and shift sizes from speech enhancement frontends, we perform signal re-synthesis before extracting features for recognition.

V. EVALUATION RESULTS

We first report speech enhancement performance and then recognition results on the CHiME-4 dataset.

TABLE II
AVERAGE SI-SDR (dB), PESQ, AND STOI (%) PERFORMANCE OF DIFFERENT METHODS ON CHANNEL 5 OF CHiME-4 (SINGLE-CHANNEL)

Methods	SI-SDR	PESQ	STOI
Unprocessed	7.5	2.18	87.0
$\hat{S}_q^{(1)}$ ($\mathcal{L}_{\text{MSA−Masking}}$)	13.9	2.94	93.9
$\hat{S}_q^{(1)}$ ($\mathcal{L}_{\text{MSA−Mapping}}$)	14.6	3.00	94.5
$\hat{S}_q^{(1)}$ ($\mathcal{L}_{\text{PSA−Masking}}$)	14.9	2.84	94.3
$\hat{S}_q^{(1)}$ ($\mathcal{L}_{\text{PSA−Mapping}}$)	15.0	2.90	94.3
$\hat{S}_q^{(1)}$ ($\mathcal{L}_{\text{RI}}$)	15.5	2.96	95.2
$\hat{S}_q^{(1)}$ ($\mathcal{L}_{\text{RI+Mag}}$)	**15.8**	**3.16**	**95.4**
SMM ($T_0^5(\lvert S_q\rvert/\lvert Y_q\rvert)$)	17.2	3.64	98.5
PSM ($T_0^1(\lvert S_q\rvert\cos(\angle S_q - \angle Y_q)/\lvert Y_q\rvert)$)	17.6	3.72	98.1

TABLE III
AVERAGE SI-SDR (dB), PESQ, AND STOI (%) OF DIFFERENT METHODS ON CHANNEL 5 OF CHiME-4 (SIX-CHANNEL)

Methods	SI-SDR	PESQ	STOI
Unprocessed	7.5	2.18	87.0
$\widehat{BF}_q^{(1)}$+*post-filtering* ($\mathcal{L}_{\text{MSA−Masking}}$)	18.6	3.32	97.3
$\widehat{BF}_q^{(1)}$+*post-filtering* ($\mathcal{L}_{\text{MSA−Mapping}}$)	19.8	3.38	97.9
$\widehat{BF}_q^{(1)}$+*post-filtering* ($\mathcal{L}_{\text{PSA−Masking}}$)	19.8	3.32	97.8
$\widehat{BF}_q^{(1)}$+*post-filtering* ($\mathcal{L}_{\text{PSA−Mapping}}$)	19.4	3.30	97.5
$\widehat{BF}_q^{(1)}$+*post-filtering* ($\mathcal{L}_{\text{RI}}$)	19.3	3.46	98.0
$\widehat{BF}_q^{(1)}$+*post-filtering* ($\mathcal{L}_{\text{RI+Mag}}$)	20.0	3.54	98.1
$\hat{S}_q^{(2)}$ ($\mathcal{L}_{\text{RI+Mag}}$)	**22.0**	**3.68**	**98.6**

TABLE IV
COMPARISON OF AVERAGE SI-SDR (dB), SDR (dB), PESQ, AND STOI (%) OF DIFFERENT APPROACHES ON CHANNEL 5 OF CHiME-4 (SIX-CHANNEL)

Methods	SI-SDR	SDR	PESQ	STOI
Unprocessed	7.5	7.6	2.18	87.0
$\hat{S}_q^{(2)}$ ($\mathcal{L}_{\text{RI+Mag}}$)	**22.0**	**22.4**	**3.68**	**98.6**
Bu *et al.* [46]	-	-	2.69	93.9
Tu *et al.* [47]	-	-	2.71	94.0
Shimada *et al.* [48]	-	16.2	2.70	94.0

TABLE V
COMPARISON OF ASR PERFORMANCE (%WER) WITH OTHER APPROACHES (SINGLE-CHANNEL)

Approaches	Dev. Set		Test Set	
	Simu.	Real	Simu.	Real
Mixtures + Trigram	8.24	6.67	12.98	10.70
+ Five-gram and RNNLM	6.58	4.84	11.17	8.38
+ LSTMLM	5.65	4.06	10.58	8.12
+ Iterative Speaker Adaptation	**4.99**	**3.54**	**9.41**	**6.82**
Kaldi baseline [44]	6.81	5.58	12.15	11.42
Du *et al.* [21]	6.61	4.55	11.81	9.15
Wang and Wang [26] (No LSTMLM)	6.77	4.99	11.14	8.28

A. Enhancement Performance

Table II compares the enhancement performance of single-channel complex-domain mapping with single-channel magnitude-domain masking and mapping, along with oracle magnitude-domain masking using the spectral magnitude mask (SMM) [1] and phase-sensitive mask (PSM) [42]. We observe better SI-SDR, PESQ and STOI results using the model trained with $\mathcal{L}_{\text{RI}}$ and $\mathcal{L}_{\text{RI+Mag}}$ than with $\mathcal{L}_{\text{MSA−Masking}}$, $\mathcal{L}_{\text{MSA−Mapping}}$, $\mathcal{L}_{\text{PSA+Masking}}$ and $\mathcal{L}_{\text{PSA+Mapping}}$, indicating the effectiveness of complex-domain estimation. Compared with $\mathcal{L}_{\text{RI}}$, $\mathcal{L}_{\text{RI+Mag}}$ yields much better PESQ (3.16 vs. 2.96), slightly better SI-SDR (15.8 vs. 15.5 dB), and marginally better STOI (95.4% vs. 95.2%). This suggests the importance of magnitude estimation for PESQ. The following experiments use $\mathcal{L}_{\text{RI+Mag}}$ as the default loss function.

Table III reports the performance of multi-channel enhancement. One straightforward approach, denoted as $\widehat{BF}_q^{(1)}$ + *post-filtering*, is to first utilize a single-channel model listed in Table II to obtain $\widehat{BF}_q^{(1)}$ via Eqs. (4)–(7) (see also Fig. 1), and then apply the single-channel model again on $\widehat{BF}_q^{(1)}$ for post-filtering. Since $\widehat{BF}_q^{(1)}$ is expected to contain low speech distortion, it can be used as the input to the single-channel model for post-filtering, although the model is trained on noisy mixtures. Clearly, using $\widehat{BF}_q^{(1)}$ + *post-filtering* obtained via the model trained with $\mathcal{L}_{\text{RI+Mag}}$ leads to the best performance. This is consistent with the single-channel results in Table II. Another approach, denoted as $\hat{S}_q^{(2)}$ (see Fig. 1), combines Y_q and $\widehat{BF}_q^{(1)}$ computed from Eqs. (4)–(7) to train another DNN for multi-channel complex spectral mapping. Clearly better results are observed over $\widehat{BF}_q^{(1)}$ + *post-filtering*, but at the expense of using one more DNN. Note that both of them show clear improvements over single-channel enhancement.

Table IV compares the proposed approach with other competitive approaches in the literature. Bu *et al.* [46] utilize estimated masks produced by BLSTM based single-channel masking to compute the signal statistics for MVDR beamforming and magnitude-domain post-filtering. Tu *et al.* [47] combine the estimated mask produced by complex Gaussian mixture models (CGMM) with the estimated ideal ratio mask (IRM) provided by an LSTM for masking-based block-wise MVDR, and use another LSTM for monaural magnitude mapping based post-filtering for further noise reduction. In [48], Shimada *et al.* combine CGMM based spatial clustering and multi-channel non-negative matrix factorization based spectral modeling to estimate time-varying speech and noise covariance matrices for time-varying beamforming. As can be observed from Table IV, substantially better enhancement results are obtained by our approach over the comparison approaches.

B. Recognition Performance

Table V reports ASR performance on the single-channel task of CHiME-4. Our single-channel system directly uses unprocessed noisy signals for recognition and obtains 6.82% WER after lattice-rescoring and iterative speaker adaptation. This result is significantly better than the previous best WERs reported by Du *et al.* [21], and Wang and Wang [26]. This result suggests that our backend is a strong one and can be very indicative at measuring the effectiveness of frontend enhancement for recognition. It should be noted that we tried to use the enhancement results of our single-channel frontends for recognition. The ASR performance is however worse than using unprocessed mixtures. This is likely due to the speech distortion introduced by DNN

TABLE VI
ASR PERFORMANCE (%WER) OF USING VARIOUS SINGLE- AND MULTI-CHANNEL MODELS FOR TI- AND TV-MVDR, AND USING TRIGRAM LANGUAGE MODEL FOR DECODING

#mics	Entry	Methods	$\hat{\Phi}^{(s)}$, $\hat{\Phi}^{(v)}$	Dev. Set		Test Set	
				Simu.	Real	Simu.	Real
2	1	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{MSA-Masking}}$)	Eq. (10), (11)	6.23	5.58	8.45	8.44
	2	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{PSA-Masking}}$)	Eq. (10), (12)	6.23	5.48	8.43	8.58
	3	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{MSA-Masking}}$)	Eq. (4), (5)	6.06	5.54	7.83	8.63
	4	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{MSA-Mapping}}$)		6.06	5.48	7.86	8.46
	5	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{PSA-Masking}}$)		6.05	5.50	7.85	8.37
	6	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{PSA-Mapping}}$)		6.15	5.50	8.18	8.36
	7	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{RI}}$)		5.98	5.52	7.82	8.23
	8	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{RI+Mag}}$)		5.93	5.48	7.68	8.29
	9	$\widehat{BF}_q^{(2)}$ ($\mathcal{L}_{\text{RI+Mag}}$)		5.91	5.42	7.74	8.11
	10	$\widehat{BF}_q^{(2)}$ ($\mathcal{L}_{\text{RI+Mag}}$)	Eq. (4), (8)	**5.32**	**5.03**	**6.85**	**7.72**
6	11	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{MSA-Masking}}$)	Eq. (10), (11)	4.16	4.24	5.16	5.75
	12	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{PSA-Masking}}$)	Eq. (10), (12)	4.04	4.15	4.87	5.55
	13	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{MSA-Masking}}$)	Eq. (4), (5)	3.98	4.24	4.75	6.08
	14	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{MSA-Mapping}}$)		3.97	4.20	4.64	5.95
	15	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{PSA-Masking}}$)		3.97	4.19	4.66	5.78
	16	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{PSA-Mapping}}$)		4.05	4.28	5.02	6.10
	17	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{RI}}$)		3.79	4.16	4.47	5.59
	18	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{RI+Mag}}$)		3.91	4.15	4.55	5.69
	19	$\widehat{BF}_q^{(1)}$ ($\mathcal{L}_{\text{RI+Mag}}$)	Eq. (13), (14)	3.86	4.11	4.51	5.49
	20	$\widehat{BF}_q^{(2)}$ ($\mathcal{L}_{\text{RI+Mag}}$)	Eq. (13), (14)	3.86	4.10	4.43	5.35
	21	$\widehat{BF}_q^{(2)}$ ($\mathcal{L}_{\text{RI+Mag}}$)	Eq. (13), (8)	**3.46**	**3.99**	**4.10**	**5.06**

TABLE VII
COMPARISON OF ASR PERFORMANCE (%WER) WITH OTHER APPROACHES (TWO-CHANNEL)

Approaches	Dev. Set		Test Set	
	Simu.	Real	Simu.	Real
$\widehat{BF}_q^{(2)}$ ($\mathcal{L}_{\text{RI+Mag}}$, Eqs. (4) and (8)) + Trigram	5.32	5.03	6.85	7.72
+ Five-gram and RNNLM	3.74	3.32	4.84	5.54
+ LSTMLM	2.52	2.15	3.28	3.80
+ Iterative Speaker Adaptation	**2.17**	**1.99**	**2.53**	**3.19**
Kaldi baseline [44]	3.94	2.85	5.03	5.40
Du *et al.* [21]	3.46	2.33	5.74	5.91

TABLE VIII
COMPARISON OF ASR PERFORMANCE (%WER) WITH OTHER APPROACHES (SIX-CHANNEL)

Approaches	Dev. Set		Test Set	
	Simu.	Real	Simu.	Real
$\widehat{BF}_q^{(2)}$ ($\mathcal{L}_{\text{RI+Mag}}$, Eqs. (13) and (8)) + Trigram	3.46	3.99	4.10	5.06
+ Five-gram and RNNLM	2.28	2.57	2.71	3.59
+ LSTMLM	1.39	1.66	1.76	2.32
+ Iterative Speaker Adaptation	**1.15**	**1.50**	**1.45**	**1.99**
Kaldi baseline [44]	2.10	1.90	2.66	2.74
Du *et al.* [21]	1.78	1.69	2.12	2.24

based enhancement and the large mismatch between the training and test conditions of CHiME-4.

Table VI presents the ASR results of TI- and TV-MVDR using $\hat{\boldsymbol{S}}^{(1)}$ and $\hat{\boldsymbol{S}}^{(2)}$ produced by the two DNNs, based on the task-standard trigram language model. We first go through the results by using the two-microphone task as an example. Entries 1–8 are obtained by using various single-channel models to compute the statistics for TI-MVDR, either by using Eqs. (4) and (5) or Eq. (10) for covariance matrix computation. Among these entries, we found that entries 7 and 8 obtain overall better WER, especially on the simulated test data. On the real test set, slightly better WER is observed. These results indicate the effectiveness of DNN based phase estimation for beamforming, especially when training and testing conditions are relatively matched. Entry 9 is obtained by using multi-channel complex spectral mapping to compute $\hat{\boldsymbol{S}}^{(2)}$, and then deriving a TI-MVDR (see Fig. 1 for more details). Slightly better WER is observed over entry 8, suggesting that the second DNN leads to better signal statistics for beamforming than the first one. Entry 10 uses $\hat{\boldsymbol{S}}^{(2)}$ to compute a TV-MVDR. Clearly better WER is observed over entry 9, indicating the effectiveness of using estimated complex spectra to compute time-varying noise covariance matrices for beamforming. On the six-microphone task, a similar trend to that in entries 3–8 is observed from entries 13–18. However, entry 18 shows slightly worse recognition performance than entry 12 (5.69% vs. 5.55% WER), likely due to the microphone failure in the six-microphone task. In entry 19, we introduce a median pooling mechanism shown in Eqs. (13) and (14) for the proposed approach to deal with microphone failure. Slight but consistent improvements are observed over entries 12 and 18 on all the four subsets. Comparing entry 20 with 19, we also observe better performance by using $\hat{\boldsymbol{S}}^{(2)}$ rather than $\hat{\boldsymbol{S}}^{(1)}$ for TI-MVDR. Using TV-MVDR in entry 21 leads to further improvement. Note that in entry 21, $\hat{\Phi}^{(v)}(f)$ in Eq. (8) is computed based on Eq. (14).

Table VII and Table VIII further apply five-gram, RNN and LSTM language models for lattice re-scoring, and perform iterative speaker adaptation for the two- and six-microphone tasks, based respectively on the TV-MVDR frontends produced in the entry 10 and entry 21 of Table VI.

Table V, Table VII and Table VIII compare the proposed system with other state-of-the-art systems. Our system advances state-of-the-art ASR results on all the tasks. The system in Du *et al.* [21] (and their journal version [22]) was the winning solution in the CHiME-4 challenge, and produces the best WER results reported to date. It ensembles one DNN and four CNN based acoustic models as the backend, using a combination of log Mel filterbank, fMLLR and i-vectors as the input features. Their frontend uses T-F masking based MVDR beamforming, where the estimated masks are combined on the basis of an unsupervised CGMM model, a supervised LSTM based IRM estimator, and frame-level voice activity detection results produced by a speech recognizer. An LSTM language model is used for lattice re-scoring. As can be seen, their frontend and backend are both ensembles of multiple models. In contrast, our system does not use any model ensemble, and obtains better ASR results on all the three tasks (6.82% vs. 9.15%, 3.19% vs. 3.91%, and 1.99% vs. 2.24% WER). These amount to 25.5%, 18.4%, and 11.2% relative WER reductions for the single-, two-, and six-microphone tasks, respectively. The improvement is especially large on the simulated test data of the two- and six-microphone tasks (2.53% vs. 5.74%, and 1.49% vs. 2.12% WER), indicating that the proposed system is particularly effective when training

and testing conditions are not very different. Another system worth mentioning is a recently-proposed CHiME-4 baseline [44] available in Kaldi. The frontend is a masking based generalized eigenvector beamformer based on a BLSTM, the acoustic model is a time-delay DNN trained with a lattice-free version of the maximum mutual information criterion, and an LSTM language model, which is the one we use in our study, is trained for lattice re-scoring. Our system obtains much better ASR results, demonstrating the effectiveness of the proposed frontend and backend.

As can be seen from entries 8 and 9, and entries 19 and 20, $\hat{\boldsymbol{S}}^{(2)}$ leads to slightly better beamforming over $\hat{\boldsymbol{S}}^{(1)}$. This is simply because $\hat{\boldsymbol{S}}^{(2)}$ is better than $\hat{\boldsymbol{S}}^{(1)}$ by using multi-channel information. As shown in Table II and Table III, $\hat{S}_q^{(2)}$ is 6.2 dB SI-SDR better than $\hat{S}_q^{(1)}$ (22.0 dB vs. 15.8 dB) in the six-channel case. Following the suggestion from an anonymous reviewer, we re-compute TI-MVDR beamformers using $\hat{\boldsymbol{S}}^{(2)}$, and then feed the beamforming results and the mixture into the second DNN to obtain $\hat{\boldsymbol{S}}^{(3)}$. $\hat{S}_q^{(3)}$ obtains 22.3 dB SI-SDR, 22.8 dB SDR, 3.69 PESQ and 98.6% STOI, which are slightly better than the $\hat{S}_q^{(2)}$ results listed in Table IV. In addition, using $\hat{\boldsymbol{S}}^{(3)}$ for TV-MVDR as is done in entries 10 and 21 leads to 7.69% and 4.95% WER on the real test set of the two- and six-microphone recognition tasks, which are slightly better than the 7.72% and 5.06% WER results.

VI. Conclusions

We have proposed a complex spectral mapping approach for single- and multi-channel speech enhancement. Experiments on CHiME-4 show that complex spectral mapping leads to better single-channel enhancement, beamforming and post-filtering, over magnitude-domain masking and mapping. Our adaptive noise covariance matrix estimation yields further ASR improvements over TI-MVDR, especially on the two-channel task. State-of-the-art results have been obtained on the enhancement and recognition tasks of the CHiME-4 corpus. Future research will consider time-domain and real-time processing, reducing the number of model parameters and extensions to multi-speaker separation.

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, pp. 1702–1726, Oct. 2018.

[2] R. Haeb-Umbach *et al.*, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 111–124, Nov. 2019.

[3] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 692–730, Apr. 2017.

[4] E. A. P. Habets and P. A. Naylor, "Dereverberation," in *Audio Source Separation and Speech Enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. New York, NY, USA: Wiley, 2018, pp. 317–343.

[5] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*, Berlin Heidelberg, Germany: Springer, 2001, pp. 157–180.

[6] 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.

[7] M. I. Mandel and J. P. Barker, "Multichannel spatial clustering for robust far-field automatic speech recognition in mismatched conditions," in *Proc. Interspeech*, 2016, pp. 1991–1995.

[8] Z.-Q. Wang, X. Zhang, and D. L. Wang, "Robust speaker localization guided by deep learning based time-frequency masking," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 178–188, Jan. 2019.

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

[10] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Comput. Speech Lang.*, vol. 46, pp. 535–557, 2017.

[11] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2016, pp. 5210–5214.

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

[13] 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. Acoustics, Speech Signal Process.*, 2017, pp. 276–280.

[14] 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. Acoustics, Speech Signal Process.*, 2017, pp. 3246–3250.

[15] 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.

[16] 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.

[17] 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. Acoustics, Speech Signal Process.*, 2019, pp. 71–75.

[18] D. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," in *Proc. IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2016, pp. 483–492.

[19] 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.

[20] K. Tan and D. L. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," in *Proc. IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2020, pp. 380–390.

[21] J. Du, Y. Tu, L. Sun, F. Ma, H. Wang, and J. Pan, "The USTC–iFlytek system for CHiME-4 Challenge," in *Proc. CHiME-4*, 2016, pp. 36–38.

[22] Y.-H. Tu *et al.*, "An iterative mask estimation approach to deep learning based multi-channel speech recognition," *Speech Commun.*, vol. 106, pp. 31–43, 2019.

[23] Z.-Q. Wang and D. L. Wang, "Unsupervised speaker adaptation of batch normalized acoustic models for robust ASR," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2017, pp. 4890–4894.

[24] Z.-Q. Wang and D. L. Wang, "Mask-weighted STFT ratios for relative transfer function estimation and its application to robust ASR," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2018, pp. 5619–5623.

[25] Z.-Q. Wang and D. L. Wang, "On spatial features for supervised speech separation and its application to beamforming and robust ASR," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2018, pp. 5709–5713.

[26] P. Wang and D. L. Wang, "Utterance-wise recurrent dropout and iterative speaker adaptation for robust monaural speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2018, pp. 4814–4818.

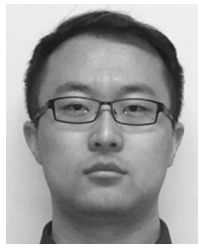
[27] S. Wisdom *et al.*, "Differentiable consistency constraints for improved deep speech enhancement," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, 2019, vol. 2019, pp. 900–904.

[28] 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.

[29] 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. Acoustics, Speech Signal Process.*, 2015, pp. 116–120.

[30] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time-frequency masks from spatial features," *Speech Commun*., vol. 68, pp. 97–106, 2015.
[31] 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.
[32] 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.
[33] 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. Acoustics, Speech Signal Process.*, 2018, pp. 5739–5743.
[34] 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. Acoustics, Speech Signal Process.*, 2018, pp. 1–5.
[35] Y. Kubo, T. Nakatani, M. Delcroix, K. Kinoshita, and S. Araki, "Mask-based MVDR beamformer for noisy multisource environments: Introduction of time-varying spatial covariance model," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2019, pp. 6855–6859.
[36] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*.
[37] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
[38] 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. 4700–4708.
[39] N. Takahashi, N. Goswami, and Y. Mitsufuji, "MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation," in *Proc. 16th Int. Workshop Acoust. Signal Enhancement, Proc.*, 2018, pp. 106–110.
[40] Y. Liu and D. L. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," in *Proc. IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2019, pp. 2092–2102.
[41] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2019, pp. 626–630.
[42] 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. Acoustics, Speech Signal Process.*, 2015, pp. 708–712.
[43] J. Heymann and R. Haeb-Umbach, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in *Proc. CHiME-4*, 2016, pp. 12–17.
[44] S.-J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," in *Proc. Interspeech*, 2018, pp. 1571–1575.
[45] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Vienna, Austria, Springer, 2014.
[46] S. Bu, Y. Zhao, M.-Y. Hwang, and S. Sun, "A robust nonlinear microphone array postfilter for noise reduction," in *Proc. IWAENC*, 2018, pp. 206–210.
[47] Y.-H. Tu, J. Du, N. Zhou, and C.-H. Lee, "Online LSTM-based iterative mask estimation for multi-channel speech enhancement and ASR," in *Proc. Annu. Summit Conf. Signal Inf. Process.*, 2018, pp. 362–366.
[48] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised speech enhancement based on multichannel NMF-informed beamforming for noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 5, pp. 960–971, May 2019.

Zhong-Qiu Wang (Student Member, IEEE) received the B.E. degree in computer science and technology from the Harbin Institute of Technology, Harbin, China, in 2013, and the M.S degree and the Ph.D. degree in computer science and engineering from The Ohio State University, Columbus, USA, in 2017 and 2020. His research interests include microphone array processing, robust automatic speech recognition, speech enhancement and speaker separation, machine learning, and deep learning.



Peidong Wang (Graduate Student Member, IEEE) received the B.E. degree in electronic information engineering from the University of Science and Technology of China, Hefei, China, in 2015. He is currently working towards the Ph.D. degree with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA. His research interests include robust automatic speech recognition, speech and language processing, microphone array processing, multimodal speech recognition, and machine learning.

DeLiang Wang (Fellow, IEEE), photograph and biography not provided at the time of publication.