Learning Complex Spectral Mapping With Gated Convolutional Recurrent Networks for Monaural Speech Enhancement

Ke Tan, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Phase is important for perceptual quality of speech. However, it seems intractable to directly estimate phase spectra through supervised learning due to their lack of spectrotemporal structure in it. Complex spectral mapping aims to estimate the real and imaginary spectrograms of clean speech from those of noisy speech, which simultaneously enhances magnitude and phase responses of speech. Inspired by multi-task learning, we propose a gated convolutional recurrent network (GCRN) for complex spectral mapping, which amounts to a causal system for monaural speech enhancement. Our experimental results suggest that the proposed GCRN substantially outperforms an existing convolutional neural network (CNN) for complex spectral mapping in terms of both objective speech intelligibility and quality. Moreover, the proposed approach yields significantly higher STOI and PESQ than magnitude spectral mapping and complex ratio masking. We also find that complex spectral mapping with the proposed GCRN provides an effective phase estimate.**

***Index Terms*—Complex spectral mapping, gated convolutional recurrent network, phase estimation, monaural speech enhancement.**

I. Introduction

SPEECH signals are distorted by background noise in daily listening environments. Such distortions severely degrade speech intelligibility and quality for human listeners, and make many speech-related tasks, such as automatic speech recognition and speaker identification, more difficult. Speech enhancement aims to remove or attenuate background noise from a speech signal. It is fundamentally challenging if the speech signal is captured by a single microphone at low signal-to-noise ratios (SNRs). This study focuses on monaural (single-channel) speech enhancement.

Monaural speech enhancement has been extensively studied in the speech processing community in the last decades. Inspired by the concept of time-frequency (T-F) masking in computational auditory scene analysis (CASA), speech enhancement has been formulated as supervised learning in recent years [36]. For supervised speech enhancement, a proper selection of the training target is important [38]. On one hand, a well-defined training target can substantially improve both speech intelligibility and quality. On the other hand, the training target should be amenable to supervised learning. Many training targets have been developed in the T-F domain, and they mainly fall into two groups. One group is masking-based targets such as the ideal ratio mask (IRM) [38], which define the time-frequency relationships between clean speech and noisy speech. Another is mapping-based targets such as the log-power spectrum (LPS) [44] and the target magnitude spectrum (TMS) [20], [12], which represent the spectral features of clean speech.

Most of these training targets operate on the magnitude spectrogram of noisy speech, which is computed from a short-time Fourier transform (STFT). Hence, typical speech enhancement systems enhance only the magnitude spectrogram and simply use the noisy phase spectrogram to resynthesize the enhanced time-domain waveform. The reason for not enhancing the phase spectrogram is two-fold. First, it was found that no clear structure exists in the phase spectrogram, which renders it intractable to directly estimate the phase spectrogram of clean speech [43]. Second, it was believed that phase enhancement is not important for speech enhancement [37]. A more recent study by Paliwal *et al.* [23], however, shows that accurate phase estimation can considerably improve both objective and subjective speech quality, especially when the analysis window for phase spectrum computation is carefully selected. Subsequently, various phase enhancement algorithms have been developed for speech separation. Mowlaee *et al.* [21] estimated the phase spectra of two sources in a mixture by minimizing the mean squared error (MSE). Krawczyk and Gerkmann [17] performed phase enhancement over voiced-speech frames while leaving unvoiced frames unaltered. Kulmer *et al.* [18] estimated the clean speech phase via the phase decomposition of the instantaneous noisy phase spectrum, followed by temporal smoothing. Objective speech quality improvements are achieved by these phase enhancement methods. Alternatively, phase information can be incorporated into T-F masking. Wang and Wang [39] trained a deep neural network (DNN) to directly reconstruct the time-domain enhanced signal using the noisy phase through an inverse Fourier transform layer. The results show that joint

Manuscript received March 21, 2019; revised August 26, 2019; accepted November 15, 2019. Date of publication November 22, 2019; date of current version December 24, 2019. This work was supported in part by the National Institute on Deafness and Other Communication Disorders under Grant R01 DC012048 and Grant R01 DC015521, and in part by Ohio Supercomputer Center. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jun Du. *(Corresponding author: Ke Tan.)*

K. Tan is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: tan.650@osu.edu).

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

Digital Object Identifier 10.1109/TASLP.2019.2955276

training of speech resynthesis and mask estimation improves perceptual quality while maintaining objective intelligibility. Another approach is the phase-sensitive mask (PSM) [5], which incorporates the phase difference between clean speech and noisy speech. The experimental results show that PSM estimation yields higher signal-to-distortion ratio (SDR) than only enhancing the magnitude spectrum.

Williamson *et al.* [43] observed that, whereas phase spectrogram lacks spectrotemporal structure, both real and imaginary components of the clean speech spectrogram exhibit clear structure and thus are amenable to supervised learning. Hence they designed the complex ideal ratio mask (cIRM), which can reconstruct clean speech from noisy speech. In their experiments, a DNN is employed to jointly estimate the real and imaginary spectra. Unlike the algorithms in [21], [17] and [18], cIRM estimation can enhance both the magnitude and phase spectra of noisy speech. The results show that complex ratio masking (cRM) yields better perceptual quality over IRM estimation while achieving slight or no improvement in objective intelligibility. Subsequently, Fu *et al.* [6] employed a convolutional neural network (CNN) to estimate the clean real and imaginary spectra from the noisy ones. The estimated real and imaginary spectra are then used to reconstruct the time-domain waveform. Their experimental results show that the CNN leads to a 3.1% improvement in short-time objective intelligibility (STOI) [29] and a 0.12 improvement in perceptual evaluation of speech quality (PESQ) [28] over a DNN. Moreover, they trained a DNN to map from the noisy LPS features to the clean ones. Their experimental results show that complex spectral mapping using a DNN yields a 2.4% STOI improvement and a 0.21 PESQ improvement over LPS spectral mapping using the same DNN.

In the last decade, supervised speech enhancement has benefited immensely from the use of CNNs and recurrent neural networks (RNNs). In [42], [41], [40] and [5], RNNs with long short-term memory (LSTM) are employed to perform speech enhancement. More recently, Chen *et al.* [1] proposed an RNN with four hidden LSTM layers to address speaker generalization of noise-independent models. They found that the RNN generalizes well to untrained speakers, and significantly outperforms a feedforward DNN in terms of STOI. In addition, CNNs have also been used for mask estimation and spectral mapping [7], [25], [11], [31]. In [25], Park *et al.* utilize a convolutional encoder-decoder network (CED) to perform spectral mapping. The CED achieves comparable denoising performance to a DNN and an RNN, while having much fewer trainable parameters. Grais *et al.* [11] proposed a similar encoder-decoder architecture. More recently, we proposed a gated residual network based on dilated convolutions, which has large receptive fields and thus can leverage long-term contexts [31]. Convolutional recurrent networks (CRNs) benefit from the feature extraction capability of CNNs and the temporal modeling capability of RNNs. Naithani *et al.* [22] devised a CRN by successively stacking convolutional layers, recurrent layers and fully connected layers. A similar CRN architecture was developed in [46]. Recently, we integrated a CED and LSTMs into a CRN, which amounts to a causal system [32]. Moreover, Takahashi *et al.* [30] developed a CRN that combines convolutional layers and recurrent layers at multiple low scales.

In a preliminary study, we recently proposed a novel CRN to perform complex spectral mapping for monaural speech enhancement [33]. This CRN was based on the architecture in [32]. Compared with the CNN in [6], the CRN yields higher STOI and PESQ, and is more computationally efficient. In this study, we further develop the CRN architecture and investigate complex spectral mapping for monaural speech enhancement. Our extensions to [33] include the following. First, each convolutional or deconvolutional layer is replaced by a corresponding gated linear unit (GLU) block [4]. Second, we add a linear layer on top of the last deconvolutional layer to predict the real and imaginary spectra.

The rest of this paper is organized as follows. In Section II, we introduce monaural speech enhancement in the STFT domain. In Section III, we describe our proposed approach in detail. Experimental setup is provided in Section IV. In Section V, we present and discuss experimental results. Section VI concludes this paper.

II. Monaural Speech Enhancement in the STFT Domain

Given a single-microphone mixture y, monaural speech enhancement aims to separate target speech s from background noise n. A noisy mixture can be modeled as

$$y[k] = s[k] + n[k], \tag{1}$$

where k is the time sample index. Taking the STFT on both sides, we obtain

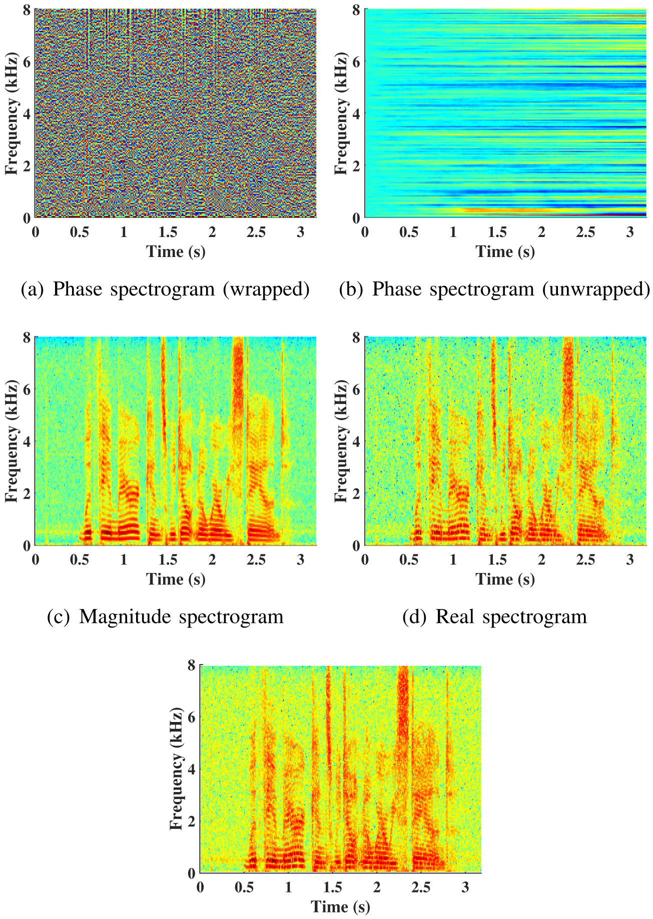
$$Y_{m,f} = S_{m,f} + N_{m,f}, \tag{2}$$

where Y, S and N represent the STFT of y, s and n, respectively, and m and f index the time frame and the frequency bin, respectively. In polar coordinates, Eq. (2) becomes

$$|Y_{m,f}|\, e^{i\theta_{Y_{m,f}}} = |S_{m,f}|\, e^{i\theta_{S_{m,f}}} + |N_{m,f}|\, e^{i\theta_{N_{m,f}}}, \tag{3}$$

where $|\cdot|$ denotes the magnitude response and $\theta_{\cdot}$ the phase response. The imaginary unit is represented by 'i'. The target magnitude spectrum (TMS) of clean speech (i.e. $|S_{m,f}|$) is a commonly-used training target in typical spectral mapping based approaches [20], [12]. In these approaches, a mapping from noisy features such as the noisy magnitude $|Y_{m,f}|$ to the target magnitude is learned. The estimated magnitude $|\hat{S}_{m,f}|$ is then combined with the noisy phase $\theta_{Y_{m,f}}$ to resynthesize the waveform. Fig. 1(a) depicts the phase spectrogram of a speech signal, where the phase values are wrapped into the range of $(-\pi, \pi]$. With the wrapping, the phase spectrogram looks rather random. An unwrapped version of the phase spectrogram leads to a smoother phase plot in Fig. 1(b), where the phase values are corrected by adding multiples of $\pm 2\pi$ when absolute phase jumps between consecutive T-F units are greater than or equal to π. One can observe that both plots exhibit no clear structure. Therefore, it would be intractable to directly estimate the phase spectrum through supervised learning.

From an alternative perspective, the STFT of a speech signal can be expressed in Cartesian coordinates. Hence, Eq. (2) can



(a) Phase spectrogram (wrapped)

(b) Phase spectrogram (unwrapped)

(c) Magnitude spectrogram

(d) Real spectrogram

(e) Imaginary spectrogram

Fig. 1. (Color Online). Illustration of phase, magnitude, real, and imaginary spectrograms of a speech signal. The magnitude, as well as the absolute values of the real and imaginary spectrograms, is plotted on a log scale.

be rewritten into

$$Y^{(r)}_{m,f} + iY^{(i)}_{m,f} = \left(S^{(r)}_{m,f} + N^{(r)}_{m,f}\right) + i\left(S^{(i)}_{m,f} + N^{(i)}_{m,f}\right), \quad (4)$$

where the superscripts (r) and (i) indicate real and imaginary components, respectively. In [43], the cIRM is defined as

$$M = \frac{Y^{(r)}S^{(r)} + Y^{(i)}S^{(i)}}{\left(Y^{(r)}\right)^2 + \left(Y^{(i)}\right)^2} + i\frac{Y^{(r)}S^{(i)} - Y^{(i)}S^{(r)}}{\left(Y^{(r)}\right)^2 + \left(Y^{(i)}\right)^2}, \quad (5)$$

where the indices m and f are omitted for simplicity. The enhanced spectrogram can be derived by applying an estimate of the cIRM $\hat{M}$ to the noisy spectrogram:

$$S = \hat{M} \times Y, \quad (6)$$

where the multiplication ' × ' above is a complex operator.

Additionally, we extend signal approximation (SA) [15]. SA performs masking by minimizing the difference between the spectral magnitude of clean speech and that of estimated speech. The loss for cRM-based signal approximation (cRM-SA) is defined as:

$$SA = \left|cRM \times Y - S\right|^2, \quad (7)$$

where $|\cdot|$ represents the complex modulus, i.e. the absolute value of a complex number.

As shown in Fig. 1(d) and 1(e), both real and imaginary spectrograms exhibit clear spectrotemporal structure, akin to the magnitude spectrogram Fig. 1(c) and thus amenable to supervised learning. Therefore, we propose to learn the spectral mapping directly from the real and imaginary spectra of noisy speech (i.e. $Y^{(r)}$ and $Y^{(i)}$) to those of clean speech (i.e. $S^{(r)}$ and $S^{(i)}$), as in [6]. Subsequently, the estimated real and imaginary spectra are combined to recover the time-domain signal.

It should be noted that Williamson *et al.* [43] claimed that directly predicting the real and imaginary components of the STFT via a DNN is not effective. However, we find that complex spectral mapping consistently outperforms magnitude spectral mapping, complex ratio masking, and complex ratio masking based signal approximation in both STOI and PESQ metrics, with a well-designed neural network architecture. For convenience, we refer to the training target used in complex spectral mapping, i.e. $S^{(r)}$ and $S^{(i)}$, as the target complex spectrum (TCS).

III. System Description

A. Convolutional Recurrent Network

In [32], we have developed a convolutional recurrent network, which is essentially an encoder-decoder architecture with LSTMs between the encoder and the decoder. Specifically, the encoder comprises five convolutional layers, and the decoder five deconvolutional layers. Between the encoder and the decoder, two LSTM layers model temporal dependencies. The encoder-decoder structure is designed in a symmetric way: the number of kernels progressively increases in the encoder and decreases in the decoder. To aggregate the context along the frequency direction, a stride of 2 is adopted along the frequency dimension in all convolutional and deconvolutional layers. In other words, the frequency dimensionality of feature maps is halved layer by layer in the encoder and doubled layer by layer in the decoder, which ensures that the output has the same shape as the input. Additionally, skip connections are utilized to concatenate the output of each encoder layer to the input of the corresponding decoder layer. In the CRN, all convolutions and deconvolutions are causal, so that the enhancement system does not use future information. Fig. 2 illustrates the CRN architecture in [32] for spectral mapping in the magnitude domain.

B. Gated Linear Units

Gating mechanisms control the information flows throughout the network, which potentially allows for modeling more sophisticated interactions. They were first developed for RNNs [14]. In a recent study [34], Van den Oord *et al.* adopted an LSTM-style gating mechanism for the convolutional modeling of images, which led to a masked convolution:

$$\begin{aligned}\mathbf{y} &= \tanh(\mathbf{x} * \mathbf{W}_1 + \mathbf{b}_1) \odot \sigma(\mathbf{x} * \mathbf{W}_2 + \mathbf{b}_2)\\ &= \tanh(\mathbf{v}_1) \odot \sigma(\mathbf{v}_2),\end{aligned} \quad (8)$$

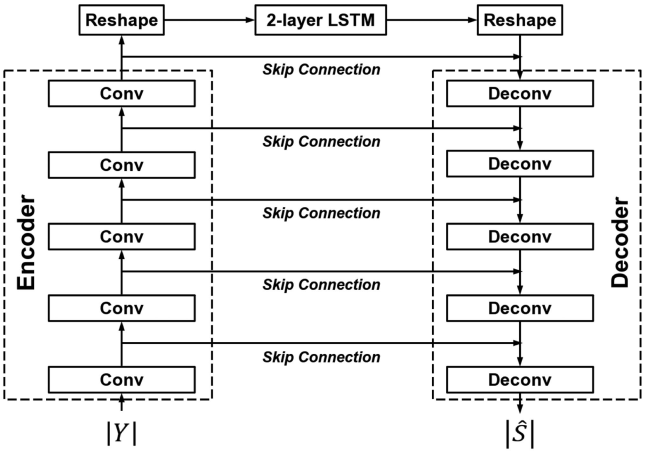


Fig. 2. Illustration of the CRN for spectral mapping in [32]. The CRN comprises three modules: An encoder module, an LSTM module, and a decoder module. 'Conv' denotes convolution and 'Deconv' deconvolution.

where $\mathbf{v}_1 = \mathbf{x} * \mathbf{W}_1 + \mathbf{b}_1$ and $\mathbf{v}_2 = \mathbf{x} * \mathbf{W}_2 + \mathbf{b}_2$. $\mathbf{W}$'s and $\mathbf{b}$'s denote kernels and biases, respectively, and σ the *sigmoid* function. Symbols $*$ and $\odot$ represent convolution operation and element-wise multiplication, respectively. The gradient of the gating is

$$\nabla[\tanh(\mathbf{v}_1) \odot \sigma(\mathbf{v}_2)] = \tanh'(\mathbf{v}_1)\nabla\mathbf{v}_1 \odot \sigma(\mathbf{v}_2) + \sigma'(\mathbf{v}_2)\nabla\mathbf{v}_2 \odot \tanh(\mathbf{v}_1), \quad (9)$$

where $\tanh'(\mathbf{v}_1), \sigma'(\mathbf{v}_2) \in (0,1)$, and the prime symbol denotes differentiation. The gradient gradually vanishes as the network depth increases due to the downscaling factors $\tanh'(\mathbf{v}_1)$ and $\sigma'(\mathbf{v}_2)$. To mitigate this problem, Dauphin *et al.* [4] introduced GLUs:

$$\begin{aligned}\mathbf{y} &= (\mathbf{x} * \mathbf{W}_1 + \mathbf{b}_1) \odot \sigma(\mathbf{x} * \mathbf{W}_2 + \mathbf{b}_2) \\ &= \mathbf{v}_1 \odot \sigma(\mathbf{v}_2). \end{aligned} \quad (10)$$

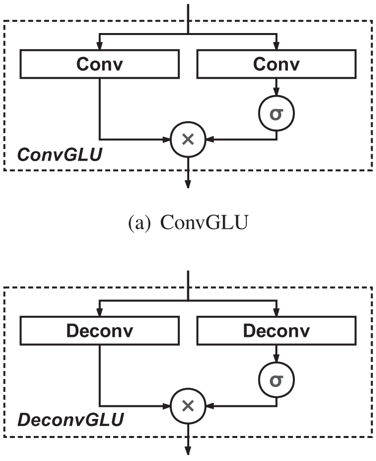
The gradient of the GLUs

$$\nabla[\mathbf{v}_1 \odot \sigma(\mathbf{v}_2)] = \nabla\mathbf{v}_1 \odot \sigma(\mathbf{v}_2) + \sigma'(\mathbf{v}_2)\nabla\mathbf{v}_2 \odot \mathbf{v}_1 \quad (11)$$

includes a path $\nabla\mathbf{v}_1 \odot \sigma(\mathbf{v}_2)$ without downscaling, which can be regarded as a multiplicative skip connection that facilitates the gradients to flow through layers. A convolutional GLU block (denoted as "ConvGLU") is illustrated in Fig. 3(a). A deconvolutional GLU block (denoted as "DeconvGLU") is analogous, except that the convolutional layers are replaced by deconvolutional layers, as shown in Fig. 3(b).

C. Model Complexity Reduction via a Grouping Strategy

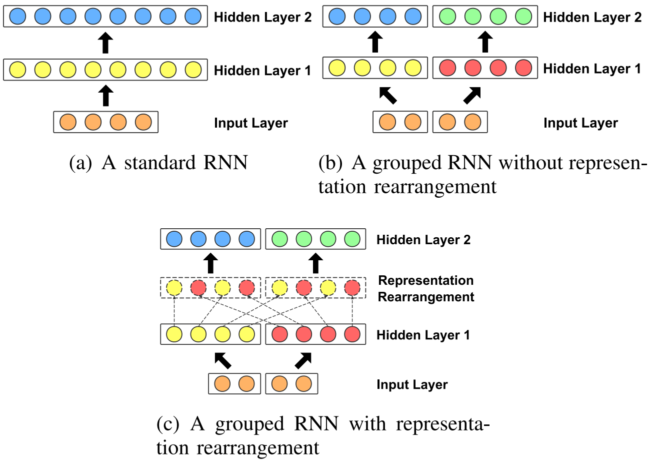
Model efficiency is important for many real-world applications. Mobile phone applications, for example, require real-time processing with low latency. In these applications, high computational efficiency and a small memory footprint are necessary. Gao *et al.* [9] have recently proposed a grouping strategy to improve the efficiency of recurrent layers while maintaining their performance. This grouping strategy is illustrated in Fig. 4. In a recurrent layer, both the input features and the hidden states are split into disjoint groups, and intra-group features are learned separately within each group, as shown in Fig. 4(b). The grouping operation substantially reduces the number of inter-layer connections and thus the model complexity. The inter-group dependency, however, cannot be captured. In other words, an output only depends on the input in the corresponding feature group, which may significantly degrade the representation power. To alleviate this problem, a parameter-free representation rearrangement layer between two successive recurrent layers is employed to rearrange the features and hidden states, so that the inter-group correlations are recovered (Fig. 4(c)). In order to elevate the model efficiency, we adopt this grouping strategy for the LSTM layers in our model. We find that this strategy improves the enhancement performance with a proper group number.



(a) ConvGLU

(b) DeconvGLU

Fig. 3. Diagrams of a convolutional GLU block and a deconvolutional GLU block, where σ denotes a sigmoid function.



(a) A standard RNN

(b) A grouped RNN without representation rearrangement

(c) A grouped RNN with representation rearrangement

Fig. 4. (Color Online). Illustration of the grouping strategy for RNNs.

D. Network Architecture

This study extends the CRN architecture in [32] (see Fig. 2) to perform complex spectral mapping. The resulting CRN additionally incorporates GLUs and thus amounts to a gated convolutional recurrent network (GCRN). Fig. 5 depicts our

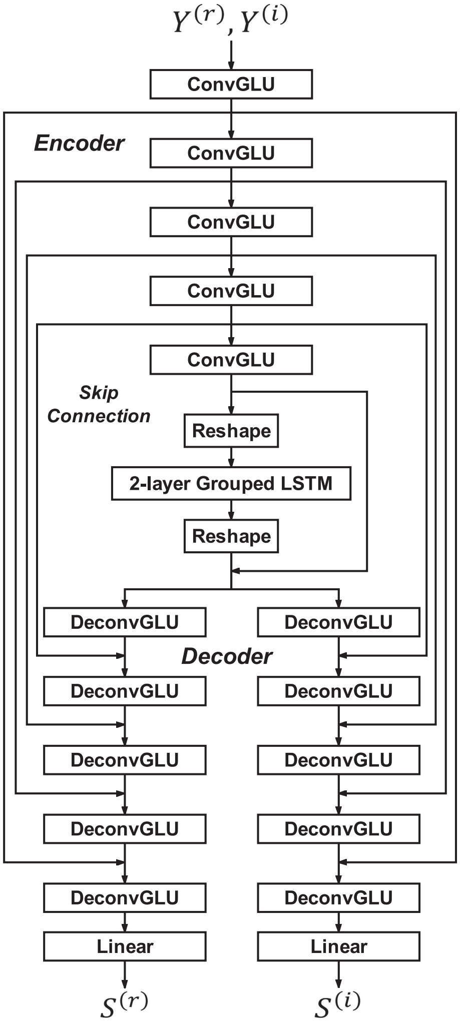


Fig. 5. Network architecture of the proposed GCRN for complex spectral mapping. More details are provided in Table I.

proposed GCRN architecture. Note that the real and imaginary spectrograms of noisy speech are treated as two different input channels as in [6]. As shown in Fig. 5, the encoder module and the LSTM module are shared across the estimates of real and imaginary components, while two distinct decoder modules are employed to estimate real and imaginary spectrograms, respectively. The design of such an architecture is inspired by multi-task learning [19], [45], in which multiple related prediction tasks are jointly learned with information shared across the tasks. For complex spectral mapping, the estimation of the real component and that of the imaginary component can be considered as two related subtasks (see [43]). Therefore, parameter sharing is expected to achieve a regularization effect between the subtasks, which may lead to better generalization. Moreover, the learning may be encouraged by parameter sharing, particularly when two subtasks are highly correlated. On the other hand, excessive parameter sharing between subtasks could discourage the learning, especially when the two subtasks are weakly correlated. Therefore, the proper choice of parameter sharing may be important for the performance. In [33], we investigated four different parameter sharing mechanisms. Among them, sharing the encoder module and the LSTM module while not sharing the decoder module leads to the best performance.

TABLE I
PROPOSED GCRN ARCHITECTURE, WHERE T DENOTES THE NUMBER OF TIME FRAMES IN THE SPECTROGRAM

layer name	input size	hyperparameters	output size
conv2d_glu_1	$2 \times T \times 161$	$1 \times 3, (1, 2), 16$	$16 \times T \times 80$
conv2d_glu_2	$16 \times T \times 80$	$1 \times 3, (1, 2), 32$	$32 \times T \times 39$
conv2d_glu_3	$32 \times T \times 39$	$1 \times 3, (1, 2), 64$	$64 \times T \times 19$
conv2d_glu_4	$64 \times T \times 19$	$1 \times 3, (1, 2), 128$	$128 \times T \times 9$
conv2d_glu_5	$128 \times T \times 9$	$1 \times 3, (1, 2), 256$	$256 \times T \times 4$
reshape_1	$256 \times T \times 4$	-	$T \times 1024$
grouped_lstm_1	$T \times 1024$	1024	$T \times 1024$
grouped_lstm_2	$T \times 1024$	1024	$T \times 1024$
reshape_2	$T \times 1024$	-	$256 \times T \times 4$
deconv2d_glu_5 (× 2)	$512 \times T \times 4$	$1 \times 3, (1, 2), 128$	$128 \times T \times 9$
deconv2d_glu_4 (× 2)	$256 \times T \times 9$	$1 \times 3, (1, 2), 64$	$64 \times T \times 19$
deconv2d_glu_3 (× 2)	$128 \times T \times 19$	$1 \times 3, (1, 2), 32$	$32 \times T \times 39$
deconv2d_glu_2 (× 2)	$64 \times T \times 39$	$1 \times 3, (1, 2), 16$	$16 \times T \times 80$
deconv2d_glu_1 (× 2)	$32 \times T \times 80$	$1 \times 3, (1, 2), 1$	$1 \times T \times 161$
linear (× 2)	$1 \times T \times 161$	161	$1 \times T \times 161$
concat	$1 \times T \times 161$ (× 2)	-	$2 \times T \times 161$

In this study, we assume that all signals are sampled at 16 kHz. A 20-ms Hamming window is utilized to produce a set of time frames, with a 50% overlap between adjacent time frames. We use 161-dimensional spectra, which corresponds to a 320-point STFT (16 kHz × 20 ms).

Table I provides details of our proposed network architecture. The input size and the output size of each layer are given in the $featureMaps \times timeSteps \times frequencyChannels$ format. In addition, the layer hyperparameters are specified in the $(kernelSize, strides, outChannels)$ format. Note that the number of feature maps in each decoder layer is doubled by skip connections. Rather than using the kernel size of 2×3 $(time \times frequency)$ in [32], we use the kernel size of 1×3, which we found does not degrade the performance. Each convolutional or deconvolutional GLU block is successively followed by a batch normalization [16] operation and an exponential linear unit (ELU) [3] activation function. A linear layer is stacked on top of each decoder to project the learned features to the real or imaginary spectrograms.

IV. EXPERIMENTAL SETUP

A. Data Preparation

In our experiments, we evaluate the proposed models on the WSJ0 SI-84 training set [26] which includes 7138 utterances from 83 speakers (42 males and 41 females). We set aside six (3 males and 3 females) of these speakers untrained for testing. In other words, we train the models with 77 remaining speakers. Of the utterances from the 77 training speakers, we hold out 150 randomly selected utterances to create a validation set with a factory noise (called "factory1") from the NOISEX-92 dataset [35] at -5 dB SNR. For training, we use 10,000 noises from a sound effect library (available at https://www.sound-ideas.com), which has the total duration of about 126 hours. For testing, we use two highly nonstationary noises, i.e. babble ("BAB") and cafeteria ("CAF"), from an Auditec CD (available at http://www.auditec.com).

Our training set contains 320,000 mixtures, and its total duration is about 500 hours. Specifically, to create a training mixture, we mix a randomly selected training utterance with a random segment from the 10,000 training noises. The SNR is randomly sampled from $\{-5, -4, -3, -2, -1, 0\}$ dB. Our test set comprises 150 mixtures that are created from 25×6 utterances of the 6 untrained speakers. We use three SNRs for the test set, i.e. -5, 0 and 5 dB.

B. Baselines and Training Methodology

We compare our proposed approach with five baselines. We first train a CRN to map from the magnitude spectrogram of noisy speech to those of clean speech [32] (denoted as "CRN + TMS"). The estimated magnitude is combined with noisy phase to resynthesize the waveform. In the second baseline (denoted as "CRN-RI + TMS"), the same CRN is employed to map from the real and imaginary spectrograms of noisy speech to the magnitude spectrogram of clean speech. Third, a CNN is trained to perform complex spectral mapping as in [6]. It has four convolutional layers with 50 kernels and the kernel size of 1×25, followed by two fully connected layers with 512 units in each layer. Parametric rectified linear units (PReLUs) [13] are employed in all layers except for the output layer. In the output layer, 322 (161×2) units with linear activations are used to predict the real and imaginary spectra. Fourth, we train our proposed GCRN to predict the cIRM. Note that the real and imaginary components of the cIRM may have a large range in $(-\infty, +\infty)$, which may complicate cIRM estimation. Therefore, we compress the cIRM with the following hyperbolic tangent as suggested in [43]:

$$O^{(x)} = K \frac{1 - e^{-C \cdot M^{(x)}}}{1 + e^{-C \cdot M^{(x)}}}, \tag{12}$$

where x denotes r or i, indicating the real and imaginary components, respectively. During inference, the estimate of the uncompressed mask can be recovered as follows:

$$\hat{M}^{(x)} = -\frac{1}{C} \log \left(\frac{K - \hat{O}^{(x)}}{K + \hat{O}^{(x)}} \right), \tag{13}$$

where $\hat{O}^{(x)}$ denotes the GCRN output. We set $K = 10$ and $C = 0.1$ as in [43]. Fifth, we train the same GCRN with the cRM-SA as the training target.

The models are trained using the AMSGrad optimizer [27] with a learning rate of 0.001. We use the mean squared error (MSE) as the objective function. The minibatch size is set to 4 at the utterance level. Within a minibatch, all training samples are padded with zeros to have the same number of time steps as the longest sample. The best models are selected by cross validation.

TABLE II
COMPARISONS OF DIFFERENT APPROACHES IN STOI AND PESQ AT -5 dB SNR

Metrics	STOI (in %)			PESQ		
Noises	BAB	CAF	Avg.	BAB	CAF	Avg.
Unprocessed	58.51	57.16	57.84	1.54	1.45	1.50
LSTM + TMS	75.65	73.15	74.40	1.94	1.94	1.94
LSTM + TCS	79.84	76.38	78.11	2.04	1.96	2.00
CRN + TMS [32]	77.10	74.49	75.80	1.99	2.00	2.00
CRN-RI + TMS	77.81	75.26	76.54	2.03	2.02	2.03
GCRN + cIRM (*G*=1)	75.97	73.53	74.75	1.95	1.93	1.94
GCRN + cRM-SA (*G*=1)	79.90	76.96	78.43	2.06	2.00	2.03
CNN + TCS (*KS*=1) [6]	67.45	67.77	67.61	1.58	1.74	1.66
CNN + TCS (*KS*=2)	67.88	69.74	68.81	1.57	1.76	1.67
CNN + TCS (*KS*=3)	69.84	70.90	70.37	1.63	1.77	1.70
CNN + TCS (*KS*=4)	69.82	71.33	70.58	1.63	1.75	1.69
CNN + TCS (*KS*=5)	70.71	71.88	71.30	1.62	1.73	1.68
GCRN + TCS (*G*=1)	**82.43**	78.98	80.71	2.15	2.06	2.11
GCRN + TCS (*G*=2)	82.42	**79.07**	**80.75**	**2.17**	2.08	**2.13**
GCRN + TCS (*G*=4)	82.20	78.72	80.46	**2.17**	2.08	**2.13**
GCRN + TCS (*G*=8)	81.46	78.29	79.88	2.15	**2.10**	**2.13**

TABLE III
COMPARISONS OF DIFFERENT APPROACHES IN STOI AND PESQ AT 0 dB SNR

Metrics	STOI (in %)			PESQ		
Noises	BAB	CAF	Avg.	BAB	CAF	Avg.
Unprocessed	70.31	69.28	69.80	1.83	1.76	1.80
LSTM + TMS	85.87	84.43	85.15	2.42	2.38	2.40
LSTM + TCS	88.85	87.44	88.15	2.54	2.46	2.50
CRN + TMS [32]	87.04	85.49	86.27	2.47	2.45	2.46
CRN-RI + TMS	87.47	86.09	86.78	2.51	2.48	2.50
GCRN + cIRM (*G*=1)	86.64	85.32	85.98	2.47	2.41	2.44
GCRN + cRM-SA (*G*=1)	89.39	88.25	88.82	2.58	2.53	2.56
CNN + TCS (*KS*=1) [6]	81.43	81.72	81.58	2.07	2.20	2.14
CNN + TCS (*KS*=2)	82.22	83.39	82.81	2.10	2.23	2.17
CNN + TCS (*KS*=3)	83.56	84.46	84.01	2.15	2.27	2.21
CNN + TCS (*KS*=4)	83.86	84.84	84.35	2.15	2.25	2.20
CNN + TCS (*KS*=5)	84.34	85.06	84.70	2.15	2.23	2.19
GCRN + TCS (*G*=1)	90.77	89.15	89.96	2.66	2.57	2.62
GCRN + TCS (*G*=2)	**90.90**	**89.34**	**90.12**	**2.70**	**2.60**	**2.65**
GCRN + TCS (*G*=4)	90.69	89.16	89.93	2.68	2.58	2.63
GCRN + TCS (*G*=8)	90.39	88.89	89.64	2.66	**2.60**	2.63

TABLE IV
COMPARISONS OF DIFFERENT APPROACHES IN STOI AND PESQ AT 5 dB SNR

Metrics	STOI (in %)			PESQ		
Noises	BAB	CAF	Avg.	BAB	CAF	Avg.
Unprocessed	81.13	80.99	81.06	2.12	2.12	2.12
LSTM + TMS	91.52	90.72	91.12	2.80	2.76	2.78
LSTM + TCS	93.03	92.35	92.69	2.90	2.84	2.87
CRN + TMS [32]	92.33	91.64	91.99	2.86	2.83	2.85
CRN-RI + TMS	92.60	91.94	92.27	2.88	2.85	2.87
GCRN + cIRM (*G*=1)	92.40	91.92	92.16	2.90	2.82	2.86
GCRN + cRM-SA (*G*=1)	94.04	93.40	93.72	2.97	2.91	2.94
CNN + TCS (*KS*=1) [6]	89.26	89.26	89.26	2.46	2.56	2.51
CNN + TCS (*KS*=2)	90.12	90.56	90.34	2.51	2.60	2.56
CNN + TCS (*KS*=3)	91.10	91.39	91.25	2.58	2.66	2.62
CNN + TCS (*KS*=4)	91.34	91.72	91.53	2.57	2.63	2.60
CNN + TCS (*KS*=5)	91.63	91.91	91.77	2.58	2.62	2.60
GCRN + TCS (*G*=1)	94.53	93.82	94.18	3.03	2.96	3.00
GCRN + TCS (*G*=2)	**94.75**	**94.02**	**94.39**	**3.07**	**2.99**	**3.03**
GCRN + TCS (*G*=4)	94.70	93.95	94.33	3.06	2.96	3.01
GCRN + TCS (*G*=8)	94.52	93.85	94.19	3.04	2.98	3.01

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Results and Comparisons

Comprehensive comparisons among different models and training targets are shown in Tables II, III and IV for − dB, 0 dB and 5 dB SNR, respectively, in terms of STOI and PESQ. The numbers represent the averages over the test samples in each test condition. The best scores in each test condition are highlighted by boldface. *KS* denotes the kernel size in the time direction and G the group number in the grouped LSTM layers. Note that $G = 1$ means that grouping is not performed. We first compare our proposed GCRN architecture with different group numbers using the TCS as the training target, as shown in the last four rows of Tables II, III and IV. It can be observed that $G = 1$, $G = 2$, $G = 4$ and $G = 8$ yield similar results in terms of both metrics, which suggests the effectiveness of the grouping strategy.

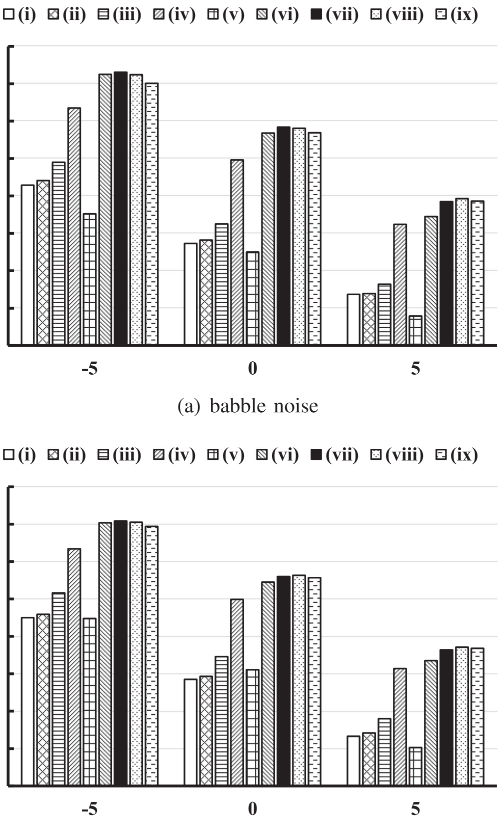
Moreover, our proposed GCRN model substantially outperforms the CNN ($KS = 1$) in [6]. At −5 dB SNR, for example, the proposed GCRN with $G = 2$ improves STOI by 13.14% and PESQ by 0.47 over the CNN. With a kernel size of 1 × 25, the CNN only captures the contexts along the frequency direction, without learning temporal dependencies. In contrast, our proposed GCRN accounts for both the frequency and temporal contexts of speech. We investigate the effects of temporal contexts for the CNN in [6] by simply using different kernel sizes in the time direction. Specifically, we use four different kernel sizes, i.e. 2 × 25, 3 × 25, 4 × 25 and 5 × 25, aside from the original version (i.e. 1 × 25) in [6]. Note that these kernels operate only on the current and past time frames, which amounts to causal convolutions. With four convolutional layers, these CNNs correspond to different temporal context window sizes of 5, 9, 13, and 17 frames, respectively. As shown in Tables II, III and IV, a larger context window size yields higher STOI but slight or no improvements in PESQ.

With the cRM-SA as the training target, our proposed GCRN yields significantly better STOI and PESQ than the same GCRN with the cIRM. Going from the cRM-SA to the TCS further improves both metrics. Take, for example, the −5 dB SNR case. The proposed GCRN ($G = 1$) with the cRM-SA yields a 3.68% STOI improvement and a 0.09 PESQ improvement compared with the estimated cIRM. An additional 2.28% STOI improvement and an additional 0.08 PESQ improvement are achieved by the estimated TCS.

We now compare spectral mapping in the magnitude domain and the complex domain. As shown in Tables II, III and IV, "CRN + TMS" and "CRN-RI + TMS" utilize the same model and training targets, but different input features. Using the real and imaginary spectra of noisy speech as the features yields slightly better STOI and PESQ than using the noisy magnitude spectra. Our proposed approach (denoted as "GCRN + TCS"), which utilizes the TCS as the training target, significantly improves STOI and PESQ over "CRN-RI + TMS". For example, "GCRN + TCS ($G = 2$)" improves STOI by 4.21% and PESQ by 0.1 over "CRN-RI + TMS" at −5 dB SNR.

To further demonstrate the effectiveness of complex spectral mapping, we additionally train two LSTM models with the TMS and the TCS, respectively. Both LSTM models have four stacked LSTM hidden layers with 1024 units in each layer, and a fully connected layer is used to estimate the TMS and the TCS, with a softplus activation function [10] and a linear activation function, respectively. As shown in Tables II, III and IV, complex spectral mapping produces consistently higher STOI and PESQ than magnitude spectral mapping.

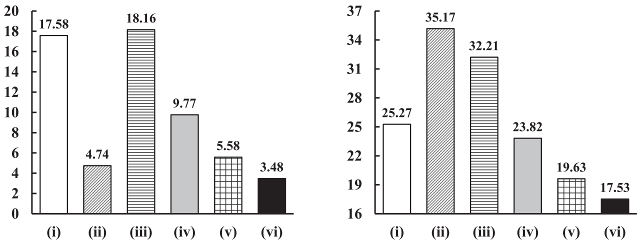
In addition, SNR improvements (ΔSNR) over the unprocessed mixtures are shown in Fig. 6. One can observe that our proposed approach produces larger SNR improvements than the baselines, with which a more than 12 dB SNR improvement is achieved at −5 dB. Fig. 7(a) shows the numbers of trainable parameters in different models, and Fig. 7(b) the numbers of floating-point fused multiply-adds that are performed to process one time frame. With the grouping strategy, our proposed model achieves higher efficiency than the CRN in [32], in terms of both computational costs and memory consumption. The CNN



(a) babble noise

(b) cafeteria noise

Fig. 6. ΔSNR in dB over the unprocessed mixtures for −5, 0 and 5 dB. The approaches are (i) CRN + TMS [32], (ii) CRN-RI + TMS, (iii) GCRN + cIRM ($G = 1$), (iv) GCRN + cRM-SA ($G = 1$), (v) CNN + TCS [6], (vi) GCRN + TCS ($G = 1$), (vii) GCRN + TCS ($G = 2$), (viii) GCRN + TCS ($G = 4$) and (ix) GCRN + TCS ($G = 8$).



(a) # trainable parameters.

(b) # floating-point fused multiply-adds.

Fig. 7. The numbers of trainable parameters (a) and floating-point fused multiply-adds per time frame (b) in different models. The unit in both parts is million. The models are (i) CRN [32], (ii) CNN [6], (iii) GCRN ($G = 1$), (iv) GCRN ($G = 2$), (v) GCRN ($G = 4$) and (vi) GCRN ($G = 8$), respectively.

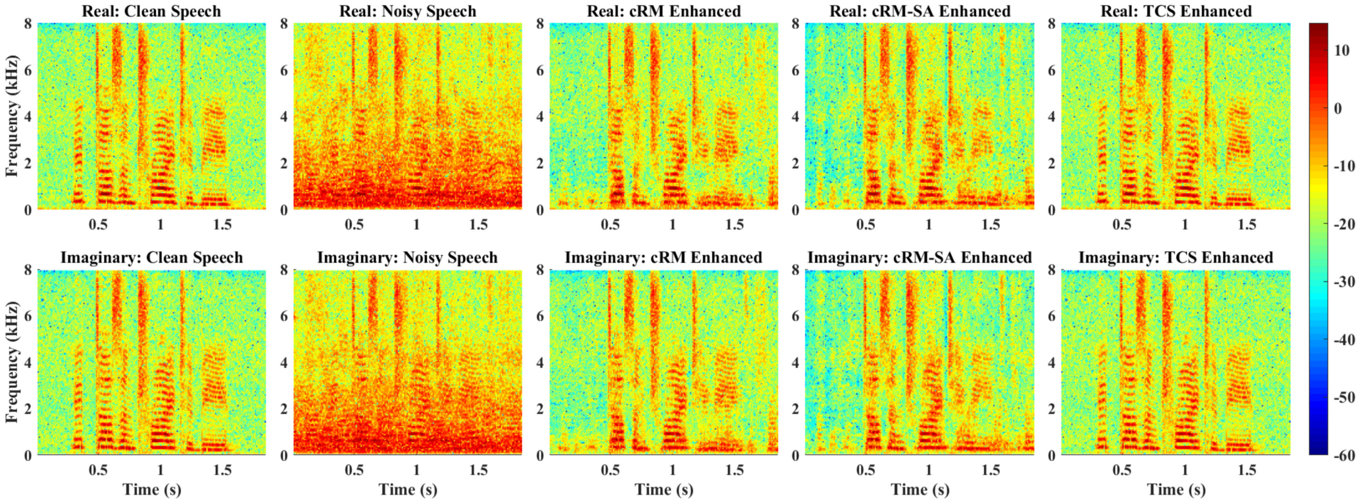


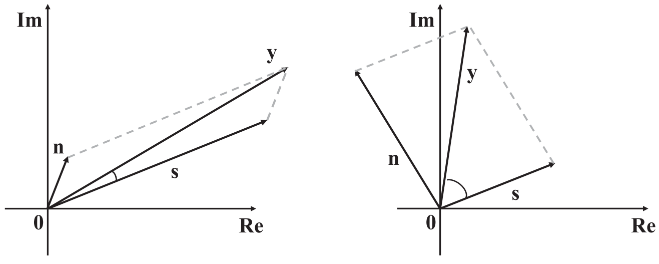
Fig. 8. (Color Online). Illustration of the real (top) and imaginary (bottom) spectrograms of clean speech, noisy speech, enhanced speech by cRM, enhanced speech by estimated cRM-SA, and enhanced speech by estimated TCS. The absolute values of the real and imaginary spectrograms are plotted on a log scale.

TABLE V

Comparisons of the Proposed Approach and Time-Domain Approaches. Here √ Indicates Causal Model, and ✗ Indicates Noncausal Model

Metrics	STOI (in %)			PESQ			Causal
SNR	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB	-
Unprocessed	57.84	69.80	81.06	1.50	1.80	2.12	-
GCRN + TCS (*G*=1)	80.71	89.96	94.18	2.11	2.62	3.00	✓
AECNN-SM [24]	79.52	90.09	94.25	2.17	2.75	3.13	✗
FCN [8]	71.08	83.67	90.46	1.65	2.18	2.57	✗
Bi-GCRN + TCS (*G*=1)	85.54	92.37	95.35	2.47	2.91	3.21	✗

in [6] has much fewer trainable parameters but higher computational costs than the CRN in [32]. With $G \geqslant 4$, our proposed GCRN has a comparable number of parameters to the CNN, but is considerably more efficient computationally. Moreover, an example of spectrograms of clean speech, noisy speech, and enhanced speech by the GCRN with the cIRM, the cRM-SA and the TCS as training targets, are shown in Fig. 8. We can see that some speech components are lost in the spectrogram of enhanced speech by the estimated cIRMor cRM-SA. In contrast, enhanced speech by the estimated TCS exhibits more similar spectrotemporal modulation patterns to clean speech and less distortion than the enhanced speech by the estimated cIRM or cRM-SA.



(a) speech-dominant (b) noise-dominant

Fig. 9. Illustration of phase error under different conditions.

We also compare our proposed approach with two recent time-domain speech enhancement approaches: AECNN-SM (autoencoder CNN with STFT magnitude loss) [24] and FCN (fully convolutional network) [8]. Additionally, we train a non-causal version of the GCRN (denoted as “Bi-GCRN”), where the LSTM layers in the middle are replaced by bidirectional LSTM layers accordingly. The comparisons are presented in Table V, in which the numbers represent the averages over the two test noises. We can see that, the GCRN improves STOI by 1.19% over the AECNN-SM at -5 dB, while the GCRN and the AECNN-SM produce similar STOI at 0 dB and 5 dB. In terms of PESQ, the AECNN-SM consistently outperforms the GCRN. It should be noted that, the AECNN-SM approach uses a much larger time frame size (i.e. 2048) than that in our approach (i.e. 320), which is likely beneficial to the AECNN-SM. It can be also observed that our approach substantially outperforms the FCN in both STOI and PESQ. Moreover, the Bi-GCRN yields significantly higher STOI and PESQ than the GCRN. This is not surprising as future frames clearly contain information useful for speech enhancement.

B. *Evaluation of Phase Estimation*

The phase of a speech signal is degraded by background noise if the noise has a different phase, as is the case in general. This is illustrated in Fig. 9(a). The degradation becomes more severe when the noise is dominant in the mixture (Fig. 9(b)). Therefore, the phase error tends to be greater in lower SNR conditions. Thus phase enhancement becomes important when the SNR is low.

TABLE VI
PHASE DISTANCES ON THE BABBLE NOISE AT DIFFERENT SNRS

SNR	$PD(S, Y)$	$PD(S, \hat{S}_{\text{cIRM}})$	$PD(S, \hat{S}_{\text{TCS}})$
-5 dB	30.896° ± 2.461°	29.516° ± 2.864°	24.151° ± 2.485°
0 dB	21.991° ± 1.641°	20.342° ± 1.549°	17.271° ± 1.363°
5 dB	14.885° ± 1.157°	13.842° ± 1.097°	12.957° ± 1.031°

TABLE VII
PHASE DISTANCES ON THE CAFETERIA NOISE AT DIFFERENT SNRS

SNR	$PD(S, Y)$	$PD(S, \hat{S}_{\text{cIRM}})$	$PD(S, \hat{S}_{\text{TCS}})$
-5 dB	38.644° ± 2.472°	36.503° ± 2.437°	30.398° ± 2.220°
0 dB	28.313° ± 1.787°	26.218° ± 1.675°	22.237° ± 1.604°
5 dB	19.867° ± 1.307°	18.455° ± 1.307°	16.613° ± 1.346°

Complex spectral mapping provides a phase estimate by solving the following two equations with two unknowns:

$$\hat{S}^{(r)} = \left|\hat{S}\right| \cos(\hat{\theta}_S) \tag{14}$$

$$\hat{S}^{(i)} = \left|\hat{S}\right| \sin(\hat{\theta}_S), \tag{15}$$

where $\hat{S}^{(r)}$ and $\hat{S}^{(i)}$ are the network outputs. To evaluate the estimated phase, we adopt two phase measures. The first is the phase distance (PD) between the target spectrogram S and the estimated spectrogram $\hat{S}$, defined in [2]:

$$PD(S, \hat{S}) = \sum_{m,f} \frac{|S_{m,f}|}{\sum_{m,'f'} |S_{m,'f'}|} \angle(S_{m,f}, \hat{S}_{m,f}), \tag{16}$$

where $\angle(S_{m,f}, \hat{S}_{m,f}) \in [0°, 180°]$ represents the angle between $S_{m,f}$ and $\hat{S}_{m,f}$. The phase distance can be regarded as a weighted average of the angle between the corresponding T-F units, where each T-F unit is weighted by the magnitude of the target spectrogram to emphasize the relative importance of the unit. The second measure quantifies the effects of the estimated phase by comparing the time-domain signals resynthesized using three kinds of phases: the noisy phase, the estimated phase and the clean phase. These phases are combined with three different magnitudes: the noisy magnitude, the enhanced magnitude by the CRN estimator in [32], and the clean magnitude.

We evaluate two estimated phases, which are calculated from the spectrogram enhanced by our proposed approach (i.e. "GCRN + TCS ($G = 1$)") and from that by "GCRN + cIRM ($G = 1$)". Tables VI and VII present the phase distance between clean and noisy spectrograms ($PD(S, Y)$) and between clean and enhanced spectrograms ($PD(S, \hat{S})$) on the babble noise and the cafeteria noise, respectively. The numbers represent the means and the standard deviations of the test samples in each test condition. One can observe that complex spectral mapping improves the phase in every condition. On the cafeteria noise at -5 dB, for example, the phase distance is improved by 8.246° on average. Moreover, "GCRN + TCS ($G = 1$)" yields consistently better phases than "GCRN + cIRM ($G = 1$)," in terms of the phase distance.

Comparisons of the signals resynthesized from the noisy phase, the estimated phase, and the clean phase are presented in Tables VIII, IX and X, respectively. As shown in Table VIII, both the objective intelligibility and the perceptual quality are improved by only enhancing the phase while keeping the noisy magnitude unaltered. For example, the phase estimated by "GCRN + TCS ($G = 1$)" improves STOI by 1.38% and PESQ by 0.12 over the noisy phase at -5 dB SNR. The clean phase yields an additional 2.05% STOI improvement and an additional 0.1 PESQ improvement at -5 dB. From Table IX, we can observe that enhancing the phase can further improve STOI and PESQ over only enhancing the magnitude, especially in low-SNR conditions (e.g. -5 dB) where phase is severely degraded. With the clean magnitude, the estimated phases improve both STOI and PESQ over the noisy phase, as shown in Table X. In addition, the phase estimated by "GCRN + TCS ($G = 1$)" produces consistently higher STOI and PESQ than the phase estimated by "GCRN + cIRM ($G = 1$)".

The above evaluations also suggest that the use of noisy phase is a significant limitation of conventional approaches that perform no phase enhancement. Our complex spectral mapping provides an effective phase estimation and avoids the use of the noisy phase.

VI. CONCLUSION

In this study, we have proposed a new framework for complex spectral mapping using a convolutional recurrent network, which learns to map from the real and imaginary spectrograms of noisy speech to those of clean speech. It provides simultaneous enhancement of magnitude and phase responses of noisy speech. Inspired by multi-task learning, the proposed approach extends a newly-developed CRN, and yields a causal, and noise- and speaker-independent algorithm for monaural speech enhancement. Our experimental results demonstrate that complex spectral mapping with our proposed model significantly improves STOI and PESQ over magnitude spectral mapping, as well as complex ratio masking and complex ratio masking based signal approximation. In addition, our proposed model substantially outperforms an existing CNN for complex spectral mapping. Moreover, we incorporate a grouping strategy into recurrent layers to substantially elevate model efficiency while maintaining the performance.

Our proposed approach also provides a phase estimate, which is demonstrated to be closer to the clean phase than the noisy phase. From another perspective, we find that the estimated phase yields noticeably higher STOI and PESQ than the noisy phase when combined with the noisy magnitude or the enhanced magnitude.

It should be noted that clean speech can be perfectly recovered from the target complex spectrogram. We believe that the GCRN-based approach with complex spectral mapping represents a significant step towards producing high-quality enhanced speech in adverse acoustic environments and practical applications. In future studies, we plan to extend our approach to multi-channel speech enhancement, in which accurate phase estimation is likely more important.

ACKNOWLEDGMENT

The authors would like to thank A. Pandey for providing his implementation of AECNN-SM for comparisons.

TABLE VIII
COMPARISONS OF NOISY PHASE, ESTIMATED PHASE, AND CLEAN PHASE COMBINED WITH NOISY MAGNITUDE IN STOI AND PESQ

Metrics	STOI (in %)									PESQ								
SNR	-5 dB			0 dB			5 dB			-5 dB			0 dB			5 dB		
Noises	BAB	CAF	Avg.	BAB	CAF	Avg.	BAB	CAF	Avg.	BAB	CAF	Avg.	BAB	CAF	Avg.	BAB	CAF	Avg.
noi_phase	58.51	57.16	57.84	70.31	69.28	69.80	81.13	80.99	81.06	1.54	1.45	1.50	1.83	1.76	1.80	2.12	2.12	2.12
est_phase$_{cIRM}$	59.01	57.21	58.11	71.10	69.64	70.37	81.83	81.44	81.64	1.58	1.50	1.54	1.87	1.82	1.85	2.16	2.17	2.17
est_phase$_{TCS}$	60.15	58.29	59.22	72.45	70.95	71.70	82.93	82.61	82.77	1.65	1.58	1.62	1.94	1.89	1.92	2.23	2.23	2.23
cle_phase	61.86	60.67	61.27	73.80	72.80	73.30	84.02	83.94	83.98	1.75	1.68	1.72	2.03	1.97	2.00	2.31	2.32	2.32

TABLE IX
COMPARISONS OF NOISY PHASE, ESTIMATED PHASE, AND CLEAN PHASE COMBINED WITH ENHANCED MAGNITUDE IN STOI AND PESQ

Metrics	STOI (in %)									PESQ								
SNR	-5 dB			0 dB			5 dB			-5 dB			0 dB			5 dB		
Noises	BAB	CAF	Avg.	BAB	CAF	Avg.	BAB	CAF	Avg.	BAB	CAF	Avg.	BAB	CAF	Avg.	BAB	CAF	Avg.
noi_phase	77.10	74.49	75.80	87.04	85.49	86.27	92.33	91.64	91.99	1.99	2.00	2.00	2.47	2.45	2.46	2.86	2.83	2.85
est_phase$_{cIRM}$	77.52	74.92	76.22	87.46	86.04	86.75	92.92	92.13	92.53	2.02	2.03	2.03	2.53	2.51	2.52	2.93	2.90	2.92
est_phase$_{TCS}$	79.31	77.18	78.25	89.25	87.91	88.58	94.05	93.31	93.68	2.15	2.14	2.15	2.68	2.63	2.66	3.07	3.01	3.04
cle_phase	81.20	79.40	80.30	90.62	89.51	90.07	95.03	94.40	94.72	2.35	2.34	2.35	2.88	2.82	2.85	3.28	3.21	3.25

TABLE X
COMPARISONS OF NOISY PHASE, ESTIMATED PHASE, AND CLEAN PHASE COMBINED WITH CLEAN MAGNITUDE IN STOI AND PESQ

Metrics	STOI (in %)									PESQ								
SNR	-5 dB			0 dB			5 dB			-5 dB			0 dB			5 dB		
Noises	BAB	CAF	Avg.	BAB	CAF	Avg.	BAB	CAF	Avg.	BAB	CAF	Avg.	BAB	CAF	Avg.	BAB	CAF	Avg.
noi_phase	93.33	93.02	93.18	95.18	94.97	95.08	96.92	96.83	96.88	2.86	2.86	2.86	3.10	3.12	3.11	3.39	3.41	3.40
est_phase$_{cIRM}$	93.77	93.48	93.63	95.96	95.73	95.85	97.60	97.49	97.55	2.92	2.97	2.95	3.24	3.28	3.26	3.53	3.58	3.56
est_phase$_{TCS}$	96.95	96.62	96.79	98.32	98.14	98.23	98.97	98.89	98.93	3.40	3.38	3.39	3.72	3.70	3.71	3.90	3.90	3.90
cle_phase	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	4.49	4.49	4.49	4.49	4.49	4.49	4.49	4.49	4.49

REFERENCES

[1] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4705–4714, 2017.

[2] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, Phase-aware speech enhancement with deep complex u-net, 2019, *arXiv:1903.03107*.

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

[4] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 933–941.

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

[6] 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 27th Int. Workshop Mach. Learn. Signal Process.*, 2017.

[7] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 3768–3772.

[8] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.

[9] F. Gao, L. Wu, L. Zhao, T. Qin, X. Cheng, and T.-Y. Liu, "Efficient sequence learning with group recurrent networks," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist.: Human Lang. Technol., Volume 1 (Long Papers)*, 2018, vol. 1, pp. 799–808.

[10] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[11] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *Proc. IEEE Global Conf. Signal Inf. Process.*, 2017, pp. 1265–1269.

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

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

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

[15] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 1562–1566.

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[17] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1931–1940, Dec. 2014.

[18] J. Kulmer and P. Mowlaee, "Phase estimation in single channel speech enhancement using phase decomposition," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 598–602, May 2015.

[19] A. Kumar and H. Daumé III, "Learning task grouping and overlap in multi-task learning," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 1383–1390.

[20] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, pp. 436–440.

[21] P. Mowlaee, R. Saeidi, and R. Martin, "Phase estimation for signal reconstruction in single-channel source separation," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 1548–1551.

[22] G. Naithani, T. Barker, G. Parascandolo, L. Bramsl, N. H. Pontoppidan, and T. Virtanen, "Low latency sound source separation using convolutional recurrent neural networks," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 71–75.

[23] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011.

[24] A. Pandey and D. L. Wang, "A new framework for supervised speech enhancement in the time domain," in *Proc. Interspeech*, 2018, pp. 1136–1140.

[25] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1993–1997.

[26] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. Workshop Speech Natural Lang.*, 1992, pp. 357–362.

[27] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[28] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 2, pp. 749–752.

[29] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[30] 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*, 2018, pp. 106–110.

[31] K. Tan, J. Chen, and D. L. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 189–198, Jan. 2019.

[32] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. 19th Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3229–3233.

[33] 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, pp. 6865–6869.

[34] A. Van Den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, *et al.*, "Conditional image generation with pixelcnn decoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4790–4798.

[35] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[36] D. L. Wang and G. J. Brown, Eds, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2006.

[37] D. L. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 4, pp. 679–681, Aug. 1982.

[38] 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.

[39] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4390–4394.

[40] F. Weninger *et al.*, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. Int. Conf. Latent Variable Anal. Signal Sepa.*, 2015, pp. 91–99.

[41] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 3709–3713.

[42] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. 2nd IEEE Global Conf. Signal Inf. Process. Symp. Mach. Learn. Appl. Speech Process.*, 2014, pp. 577–581.

[43] D. S. 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, 2016.

[44] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[45] Y. Zhang and Q. Yang, "An overview of multi-task learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 30–43, 2018.

[46] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2401–2405.

Ke Tan 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 toward the Ph.D. degree with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA. His research interests include speech enhancement and separation, microphone array processing, robust speech recognition, and deep learning.

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