



A Multi-Scale Feature Recalibration Network for End-to-End Single Channel Speech Enhancement

Yang Xian , *Student Member, IEEE*, Yang Sun , *Member, IEEE*, Wenwu Wang , *Senior Member, IEEE*, and Syed Mohsen Naqvi , *Senior Member, IEEE*

Abstract—Deep neural networks based methods dominate recent development in single channel speech enhancement. In this paper, we propose a multi-scale feature recalibration convolutional encoder-decoder with bidirectional gated recurrent unit (BGRU) architecture for end-to-end speech enhancement. More specifically, multi-scale recalibration 2-D convolutional layers are used to extract local and contextual features from the signal. In addition, a gating mechanism is used in the recalibration network to control the information flow among the layers, which enables the scaled features to be weighted in order to retain speech and suppress noise. The fully connected layer (FC) is then employed to compress the output of the multi-scale 2-D convolutional layer with a small number of neurons, thus capturing the global information and improving parameter efficiency. The BGRU layers employ forward and backward GRUs, which contain the reset, update, and output gates, to exploit the interdependency among the past, current and future frames to improve predictions. The experimental results confirm that the proposed MCGN method outperforms several state-of-the-art methods.

Index Terms—Bidirectional gated recurrent unit (BGRU), feature recalibration, multi-scale convolutional layer, single channel, speech enhancement.

I. INTRODUCTION

THE intelligibility and quality of the speech signal recorded in a real acoustic scene are often degraded by the background noise and interfering sound in the environment. Speech enhancement aims to recover the target speech by removing the background noise and interfering sound from noisy speech mixtures. Single channel speech enhancement refers to the scenario, where only a single mixture is available, which is an extreme case of the under-determined problem, i.e. the number of sources is greater than the number of mixtures. Such a problem can be found in many real-world applications, such

Manuscript received April 20, 2020; revised August 24, 2020, November 5, 2020, and December 10, 2020; accepted December 11, 2020. Date of publication December 18, 2020; date of current version January 29, 2021. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Pavel Rajmic. (*Corresponding author: Yang Xian.*)

Yang Xian and Syed Mohsen Naqvi are with the Intelligent Sensing and Communications Research Group, School of Engineering, Newcastle University, Newcastle, Tyne NE1 7RU, U.K. (e-mail: Y.Xian2@newcastle.ac.uk; Mohsen.Naqvi@newcastle.ac.uk).

Yang Sun is with the Big Data Institute, University of Oxford, Oxford OX3 7LF, U.K. (e-mail: Yang.sun@bdi.ox.ac.uk).

Wenwu Wang is with the Center for Vision Speech and Signal Processing, Department of Electrical and Electronic Engineering, University of Surrey, Surrey GU2 7XH, U.K. (e-mail: W.Wang@surrey.ac.uk).

Digital Object Identifier 10.1109/JSTSP.2020.3045846

as mobile communication, automatic speech recognition, and robotics [1]–[5].

A wide variety of methods have been proposed for speech enhancement. Conventional methods include statistical methods, such as Wiener filtering [6] and minimum mean-square error (MMSE) estimation [7], based on statistical modelling of spatial, spectral, or temporal features derived from the sensor signals. For instance, the MMSE estimator achieves speech enhancement by modeling the speech and noise spectral components as statistically independent Gaussian random variables.

Deep neural networks (DNNs) are today considered state-of-the-art in speech enhancement. Unlike the conventional methods, the DNNs based methods [8]–[10] aim to learn a mapping or masking relationship between the representations of noisy speech mixture and target speech, via a training process. Then, the trained model is used to make prediction of the target speech directly (via mapping) [11], or the T-F mask (via masking) [12]–[14], where either ideal binary mask (IBM) or ideal ratio mask (IRM) has been used as the training target. Recent results show that the mapping based methods outperform the masking based methods [15].

Different from vanilla DNNs, recurrent neural network (RNN) has been used for temporal modelling of speech and offers advantages in mismatched conditions [16]. In particular, long short-term memory (LSTM) [17] employs the cell memory, input, output and forget gates to capture the interdependency between the past and current frames, which improves the accuracy of the estimation for the mask and mapping relations [18]. Previous results show that it improves enhancement performance in the case of unseen speakers [15], [16]. As an extension to LSTM, the bidirectional LSTM has been proposed to also considers the impact of the future frames, thus capturing the long-term interdependency among the past, current and future frames [15].

Another promising direction has been on the exploitation of convolutional neural network (CNN), such as [19], where a convolutional encoder decoder (CED) is introduced to estimate the mapping relation between the noisy mixture and target speech. This is further improved for learning multi-resolution features, with a multi-resolution convolutional auto-encoders (MCARE) model [20], learning with dilated convolution to enlarge the receptive fields of the network in Wavenet, and learning with a gated mechanism to control the information flow among each layer [21]. Furthermore, the gated recurrent network (GRN) method is used with dilated 2-D convolutional layers to enlarge the receptive fields in the time-frequency (T-F) domain [15].

The recurrent and convolutional architectures have been used together to further improve enhancement performance. For example, in the convolutional recurrent network (CRN) [22], the convolutional encoder-decoder is integrated with the LSTM, where the CED is used to capture the local T-F patterns, and the LSTM is used to capture long-term interdependency [22]. The CRN method was shown to perform better than the LSTM.

All the above methods are supervised methods where class labels are required for training the model. In contrast, unsupervised methods have also been proposed for speech enhancement without the requirement of class labels. A well-known method is the speech enhancement generative adversarial network (SEGAN) method [23].

The aforementioned methods are promising and represent current state-of-the-art. However, there are still several limitations. For the CED and CRN methods, a fixed kernel (filter) size is often used. The local information (i.e. feature) in the signal can be extracted by using a kernel of small size, while the contextual feature needs to be extracted with a larger kernel size. A method that can extract both local and contextual information is desired. In the LSTM and CRN models, causal systems are often designed by considering only current and past samples from the signal. However, in terms of [21], the prediction performance of the model can be further improved by considering the future samples. Therefore, in our work, the future information (i.e. a non-causal system) is considered to improve the enhancement performance.

In addition, the implementation of LSTM often involves computational loads for calculating the input, output, forget gates and cell memory [17], [24], sometimes, this can be problematic when the models are deployed on resource-limited devices. It would be desirable to use more efficient RNN models such as GRU/BGRU, with performance comparable to LSTM/BLSTM but less memory requirements. In addition, in the Inception network [25], the features of different scales are concatenated directly, and they are assigned with the equal weight. This means that features are considered as equally important, which may be problematic especially when the features are induced by noise. This could be further improved by assigning features with different weights, as shown in our work.

In this paper, we propose a multi-scale feature recalibration convolutional bidirectional GRU network (MCGN), with following specific contributions.

First, we introduce a multi-scale feature recalibration (MCFR) convolutional encoder-decoder module, where the kernels with different sizes are exploited in each convolutional layer, to obtain features in different scales. This helps capture the interdependency between the local and contextual information within the signal, and allows the feature in each scale to be assigned with a different weight in order to retain the components from speech while suppressing the components from noise.

Second, the bottleneck convolutional layers are introduced, which uses the 1-D convolutional layer with kernels of size (1,1) to compress the information flow inside the proposed MCGN.

Third, connection layers are used in MCGN, including fully connected (FC) layer and BGRU layers. The FC layer is exploited to reduce the dimension of encoder output. The BGRU layers can capture the inter-dependencies among the past,

current and future temporal frames. Compared with BLSTM, they offer similar performance but require fewer parameters.

Fourth, the multi-scale convolutional output layer is proposed to accelerate the convergence. The output layer enables the enhanced output with access to the different scale convolutional operators, which facilitate network training.

The remainder of the paper is organized as follows. Section II describes the proposed MCGN method. The experimental settings and results are discussed in Section III. Section IV states the conclusions.

II. PROPOSED METHOD

A. Problem Statement

In single channel speech enhancement, the noisy speech mixture can be written as:

$$y(m) = s(m) + n(m) \quad (1)$$

where $y(m)$ denotes the noisy speech, $s(m)$ and $n(m)$ represent the clean speech signal and noise at discrete time m , respectively. By using the short-time Fourier Transform (STFT), the noisy speech mixture at time frame $t \in [1, 2, \dots, T-1, T]$ and frequency bin $f \in [1, 2, \dots, F-1, F]$ is represented as:

$$Y_{t,f} = S_{t,f} + N_{t,f} \quad (2)$$

where $S_{t,f}$ and $N_{t,f}$ are the STFT of the clean speech signal and noise, respectively. The neural network model is trained to find the mapping relation G_θ between the magnitude spectrum of the clean speech signal $|S_{t,f}|$ and the noisy speech mixture $|Y_{t,f}|$, G_θ is parametrized by θ . The mapping function is estimated by optimizing the loss function as:

$$\begin{aligned} Loss &= \frac{1}{TF} \sum_{t=1}^T \sum_{f=1}^F [G_\theta(|Y_{t,f}|) - |S_{t,f}|]^2 \\ &= \frac{1}{TF} \sum_{t=1}^T \sum_{f=1}^F (|\hat{S}_{t,f}| - |S_{t,f}|)^2 \end{aligned} \quad (3)$$

where $|\hat{S}_{t,f}|$ is the magnitude spectrum of the estimated target speech, which is combined with phase information of the noisy mixture to recover the target speech.

B. Proposed Network Architecture

The details of the proposed MCGN architecture are shown in Fig. 1. The MCGN contains four parts, i.e. convolutional encoder, convolutional decoder, connection layers, and multi-scale convolutional output layers. The magnitude spectrum of the noisy mixture is fed to the proposed MCGN, which outputs the estimated magnitude spectrum of the target speech. The convolutional encoder consists of six convolutional layers containing four multi-scale convolutional layers, an input convolutional (the first) layer and a bottleneck convolutional layer. The multi-scale convolutional layers contain five sub convolutional blocks with varied kernel sizes. Similarly, the convolutional decoder has a symmetric structure with the convolutional encoder. The output of the convolutional encoder is fed to the connection layers. After processed by the connection layers, the information flow is fed

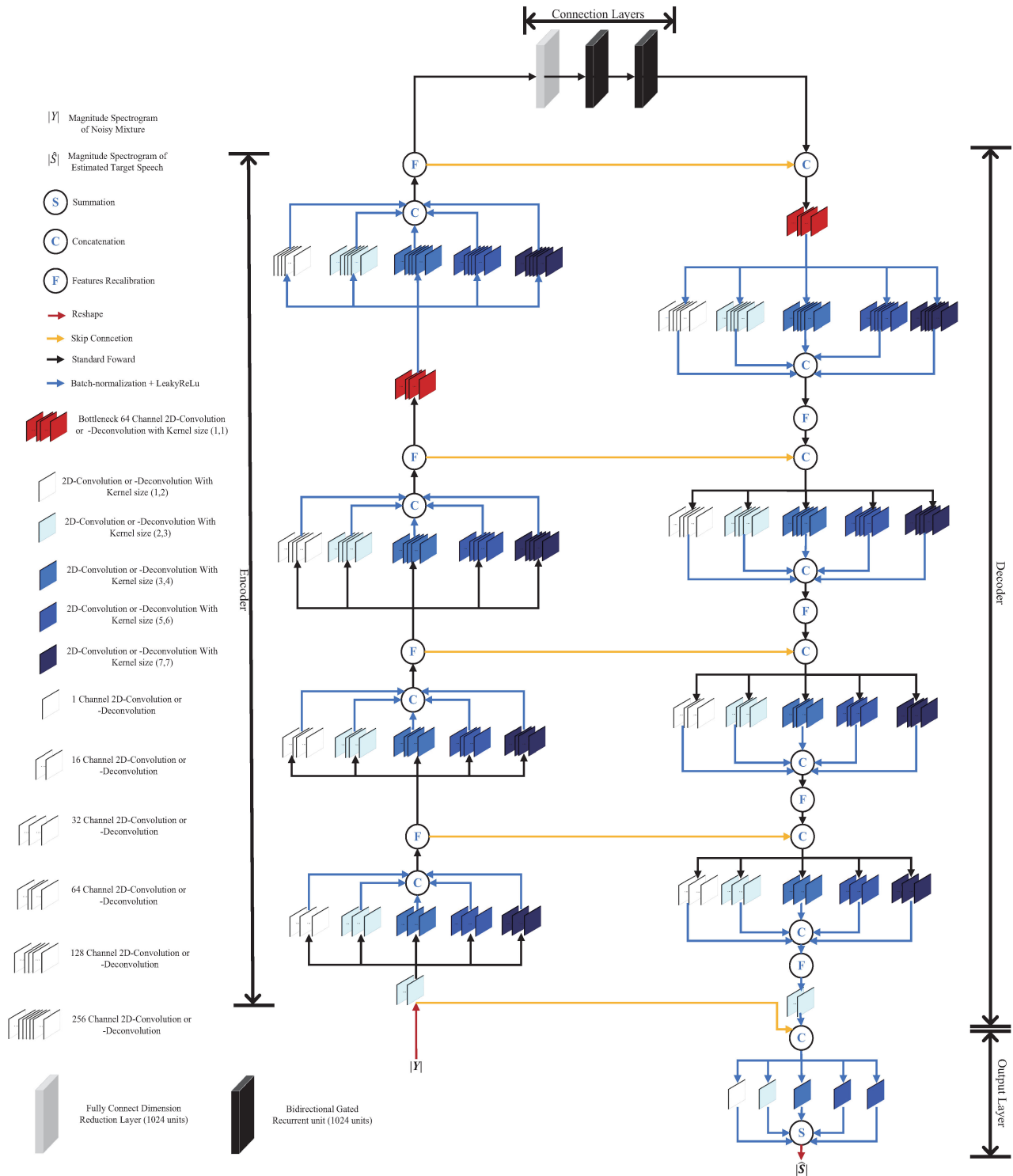


Fig. 1. Architecture of the proposed MCGN. The components and their functions are shown at the left of the figure. The overlapped 3-D boxes represent the multi-channel 2-D convolutional neural networks. The colored arrows and named circles represent the information flow and operations. The convolutional encoder is on the left of the figure, and the convolutional decoder is on the right of the figure, connection layers are shown in the middle of figure. The figure is color-coded to facilitate understanding.

to the convolutional decoder. In addition, the skip connections are added among the convolutional encoder and decoder. The layer hyper-parameters can be found in Fig. 1. The stride size of all layers is (1,2), except the multi-scale output layer, which has a fixed stride size (1,1).

C. Multi-Scale Feature Recalibration Convolutional Layer

The receptive field is a region where CNN can affect a particular high-level feature. A small receptive field is feasible to extract local information, and a large receptive field offers

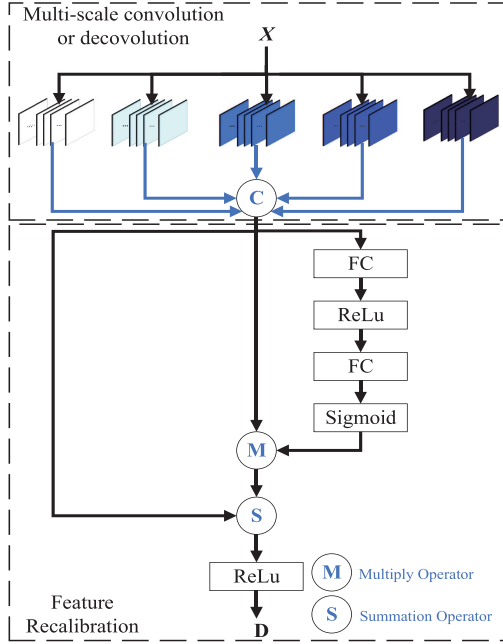


Fig. 2. Multi-scale feature recalibration network, where X , and D represent the input and output of the MCFR module, respectively. The multi-scale convolution or deconvolution is shown on top of the figure, the bottom of the figure shows the feature recalibration module.

contextual information [15]. In conventional CNN, a fixed kernel size is often used, as a result, it compromises between local and contextual information extracted from the signal. To address this limitation, a multi-scale convolutional feature recalibration (MCFR) layer is designed to capture the information on different scales and generate the multi-scaled feature. As shown in Fig. 2, MCFR contains several convolutional operators, which use the kernels of different sizes to capture the information with various scales. The convolutional operators with the small kernel sizes can extract the feature from the short duration speech, thus capturing the adjacent T-F points local dependency. The smallest kernel size (1,2) is employed, which allows the feature from two adjacent T-F points to be extracted. The convolutional operators with large kernel sizes offer large receptive fields and can extract features from long-duration speech. These features contain contextual information compared with the feature extracted by kernels with smaller sizes. The batch-normalization is used after each convolutional operator. Different from the standard CNN, which uses the ReLU activation function [26], our proposed MCGN utilizes the activation function LeakyReLU [27]. Then, we concatenate the outputs of each convolutional operator into a single output vector, forming the input of the next stage, as shown in Fig. 2. The multi-scale deconvolutional layer has a similar structure as the one in MCFR, by replacing the convolutional operators with deconvolutional operators.

After the features at different scales are extracted by using the convolutional operators with varied kernel sizes, a feature recalibration module is introduced to help the network to be selective when using these scaled features, i.e. by assigning different weights to features. It is shown on the bottom of Fig. 2. We refer to the proposed multi-scale convolutional feature recalibration layer as the MCFR layer. In the MCFR layer, we use

n sub-convolutional blocks, and each block has the same number of channels but different kernel sizes to capture the features in different scales. The input of the multi-scale layer is X , and the output is $K = [k_1, k_2, \dots, k_n]$, where k_n is captured by the n -th sub 2-D convolutional block that has different kernel size compared with other 2-D convolutional blocks.

There are several operations for estimating the recalibration coefficients, based on two criteria: the recalibration coefficient could capture the nonlinear relation inside the multi-scaled feature, and allocate relatively higher weights to speech components and lower weights to noise components within the feature. We use the following operations to meet these criteria: two FC layers, ReLU and Sigmoid activations. These operations are shown as follows,

$$c_{1n} = w_{1n} \odot k_n + b_{1n} \quad (4)$$

$$a_n = \max[0, c_{1n}] \quad (5)$$

$$c_{2n} = w_{2n} \odot a_n + b_{2n} \quad (6)$$

$$rs_n = \frac{e^{c_{2n}}}{e^{c_{2n}} + j} \quad (7)$$

where w_{1n} , w_{2n} denote the weight parameters, \odot denotes element-wise multiplication, b_{1n} , b_{2n} represent the biases. c_{1n} and c_{2n} represent the operations in FC1 and FC2 layers, respectively. $j = [1, 1, \dots, 1]$, and it has the same dimension as c_{2n} . The exponential function e is operated element-wise on c_{2n} , so is the division in the right hand side of equation (7). The vector rs_n contains the recalibration coefficient of the n -th scaled feature. Empirically, we opt for the ReLU function as (5), which is employed as a non-negative constraint. Inspired by the success of the gating mechanism, we introduce Sigmoid as a gating function to control the information flow, which aims to assign different weights to speech and noise components. The rescaled n -th feature is:

$$p_n = k_n \odot rs_n \quad (8)$$

Therefore, the rescaled multi-scale feature is $P = [p_1, p_2, \dots, p_n]$. We introduce deep skip connection (as in residual learning [28]) inside the MCFR layer. In addition, the residual learning does not introduce any additional parameters. Mathematically, the original relation for the MCFR layer is $D = P$, by using the residual learning and the ReLU function, the relation becomes:

$$D = \max[0, K + P] \quad (9)$$

Following the extraction of multi-scale features, the proposed MCGN learns the weights and applies them to these features which help retain speech components and suppresses the noise components in the noisy mixture.

D. Bottlenecks Convolutional Layers

One of the practical problems in multi-scale convolutional layers that need to be solved is the concatenation of the multi-scale features, which would increase the dimension of the features and cause an increase in computational cost. Therefore, a structure that can retain the information while reducing the complexity (e.g. dimension) is needed. Inspired by the embedding

techniques that a low dimensional embedding might contain sufficient information about a relatively large patch [25], [29], we introduce the bottleneck convolutional layers in the proposed MCGN architecture. The bottleneck convolutional layer is a 2-D convolutional layer with (1,1) kernels and 64 channels, followed by the batch-normalization and LeakeyReLU [27]. It is located before the last convolutional encoder layer and the first decoder layer, as shown in Fig. 1 (red convolutional blocks). The first bottleneck convolutional layer reduces the dimension from 640-D to 64-D for the last encoder layer, and the second bottleneck convolutional layer reduces the dimension from 128-D to 64-D for the first decoder layer.

E. Connection Layers

The original convolutional encoder-decoder does not well utilized the long-term temporal information, which, nevertheless, may be valuable in speech enhancement [16], [22]. The CRN method uses the LSTM to capture the long-term interdependency between the past and current temporal frames. However, CRN is designed for the casual problem, which utilizes long-term interdependency between past and current temporal frames. According to [21], the future frames could be used to improve enhancement performance. In our work, we introduce BGRU to capture the long-term interdependency among the past, current and future temporal frames. In comparison, GRU offers comparable performance to LSTM [24], [30], [31], but has an advantage in parameter efficiency. However, the merging of the multi-scaled convolutional sub-blocks would lead to an inevitable increase in its dimension. Therefore, it is necessary to find a way to retain the information and, at the same time, to reduce the dimension and computational cost. To address this, we use a fully connected (FC) layer, as the number of parameters of the fully connected dense layer is smaller than that of the RNN based layer, leading to a reduced dimension in the output of the FC layer, as compared with the output of the encoder.

F. Multi-Scale Output Layer

We add the skip connection from the input to the multi-scale output layer, as shown at the bottom of Fig. 1. As a result, the multi-scale output layer can estimate the magnitude of the target speech from the previous layer's information flow and the input magnitude of the noisy mixture. The multi-scale output layer is a 2-D deconvolutional layer, which contains five sub-blocks, and the kernel sizes of these sub-layers are different. Unlike the MCFR layer, these varying scaled features are concatenated, the different scaled features are summed together to generate an output matrix with the same size as the input matrix. Thus, the multi-scale output layer utilizes local and contextual information. The stride size of the output layer is set to (1,1). Batch-normalization and linear activation are followed.

III. EXPERIMENTAL EVALUATIONS

A. Datasets

We evaluate our system with three experiments using three different datasets. In the first experiment, we use 1000 clean utterances mixed with 20 noise signals to generate the training

set in our first experiment. The clean utterances are randomly selected from the TIMIT corpus [32], and noise files are selected from Non-Speech Sounds [33] and NOISEX-92 [34] datasets. Similarly, 100 clean utterances are mixed with 6 noise signals to generate the testing datasets. To better evaluate enhancement performance, the speakers in the training set are different from the speakers in the testing dataset. Meanwhile, the testing noisy interferences are categorized into two types, the seen noises (Babble, Leopard, F16) and the unseen noises (N56, N72, White). Babble, Leopard, F16, N56, N72 are non-stationary noises, and White is stationary noise. N56 and N72 are wind and water sounds, respectively. The noisy mixtures are generated by mixing the clean utterances and noises at -5 dB, 0 dB and 5 dB signal-to-noise ratio (SNR) levels. In total, about 50 hours ($3 \times 3 \times 1000 \times 20 \div 3600$) noisy mixtures are used to train the networks.

In the second experiment, we evaluate the proposed method on a published dataset [21], [23]. The datasets are generated by using the VCTK corpus [35] and Environment Multichannel Acoustic Noise Database [36]. The utterances from 28 speakers and 2 speakers are used for training and testing, respectively. Each speaker has spoken around 400 sentences. The training utterances are mixed with 10 types of noise in four SNR levels (0 dB, 5 dB, 10 dB and 15 dB). In total, there are 11 572 noisy mixtures for training. Similarly, the testing utterances are mixed with 5 types of noise in four SNR levels (2.5 dB, 7.5 dB, 12.5 dB and 17.5 dB). In total, the testing set includes 824 noisy mixtures, where both the speakers and noises are unseen in the training set.

In the third experiment, we evaluate the proposed MCGN method with a larger dataset. For the training set, we randomly select 2500 clean utterances from the TIMIT [32] and VCTK [35] corpora, mix them with 20 different noise signals selected from the Non-Speech Sounds [33] and NOISEX-92 [34] datasets, to generate 50 000 training mixtures for each SNR level (-5 dB, 0 dB, and 5 dB). Similarly, for the testing set, we randomly select 500 clean utterances and mix them with 5 different noise signals, to generate 2500 noisy mixtures for each SNR level. The speakers of the training dataset are different from those in the testing dataset. The Babble, Leopard, F16 are seen noises, while N56 and N72 are unseen noises.

The signal to distortion ratio improvement (Δ SDR) [37], perceptual evaluation of speech quality (PESQ) [38] and short-time objective intelligibility (STOI) [39] are used to measure the performance. The Δ SDR is equal to the SDR of the estimated speech minus the SDR of the unprocessed noisy mixture. The PESQ ranges from -0.5 to 4.5, which indicates the speech perception quality score. The STOI ranges from zero to one, which indicates the intelligibility quality of human speech. The higher values of the measurements indicate better enhancement performance.

B. Baselines and Parameters

The proposed MCGN is compared with seven baseline methods, including the standard DNN method from [11], the DNN method with skip connection S-DNN from [10], the LSTM model used in [16], the BLSTM model used in [15], the CNN

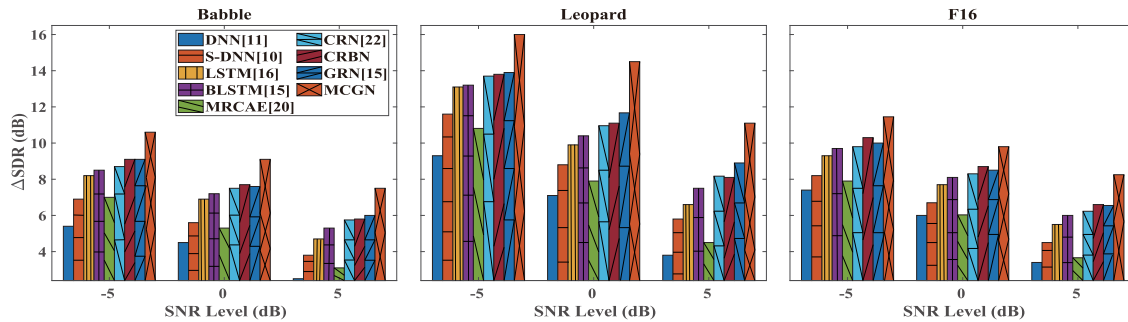


Fig. 3. Speech enhancement performance comparison in terms of Δ SDR for three types of seen noises with different methods and SNR levels. Each result is the averaged value of 100 independent experiments.

based methods, the MRCAE method from [20], and the GRN method in [15]. The parameters of the CRN model are set by following [22]. LSTM and BLSTM have four hidden layers, where each hidden layer contains 1024 units with a dropout rate of 0.2, and the output layer is a dense layer. The MRCAE is a five-layered 1-D convolutional encoder decoder. The encoder consists of two multi-resolution 1-D convolutional layers, and the decoder mirrors the encoder. A deconvolutional layer is used as the output layer of MRCAE. The CRN consists of the 2-D convolutional encoder, two-layered LSTM and 2-D convolutional decoder, which are connected by standard feed-forward connections and skip connections. The GRN is a 62-layered fully connected dilated convolutional neural network with the residual. The aforementioned baseline methods and proposed MCGN method take the STFT magnitude spectrum of the noisy speech mixture as the input features, and output the corresponding magnitude spectrum of the estimated target speech. The estimated magnitude spectrum is combined with the noisy phase to re-synthesize the estimated target speech waveform. Furthermore, the proposed MCGN model trained on the published dataset [21], [23] is compared with the SEGAN and Wavenet. The SEGAN employs generator and discriminator to learn and judge the input data distribution, which uses the adversarial training [23]. The Wavenet is a 30-layered fully connected convolutional neural network [21].

The input and output layers for all methods contain 257 units. The baseline methods and proposed MCGN method are trained with the Adam optimization algorithm [40]. The initial learning rate is set to 0.0001. The mean square error (MSE) is employed as the objective function for the baseline and the proposed MCGN methods. The dropout rate is fixed to 0.2. The sample rate of noisy speech mixtures is 16 kHz, and the window length is 512. The time resolution is 32 ms, and frequency resolution is 32.15 Hz. The next two sections (i.e. Sections III.C and III.D) report the results based on the first dataset, while Section III.E and III.F present results for the second and third dataset, respectively.

C. Unseen Speakers With Seen Noises

Fig. 3 and Table I provide experimental results in terms of Δ SDR, STOI and PESQ for the baseline and the proposed methods with real-world noises. The speakers used in testing are

unseen in the training data. The noises used in testing include Babble, Leopard, and F16.

The DNN generates, on average, Δ SDR = 5.49 dB, STOI = 76.26% and PESQ = 2.07, which offers the worst enhancement performance across all the compared methods. These results show that effectiveness of DNN remains insufficient. The S-DNN slightly outperforms the DNN, because S-DNN explicates the skip connection. The MRCAE method uses the multi-resolution 1-D convolutional encoder decoder and offers a small improvement over the DNN in terms of Δ SDR, and PESQ.

The LSTM generates, on average, Δ SDR = 8.03 dB, STOI = 78.77% and PESQ = 2.33, which shows advantages over the DNN, S-DNN and MRCAE. Unlike the DNN, S-DNN and MRCAE method, the LSTM exploits the memory cell to keep the hidden states from the past temporal frame. The interdependency between them are captured by the LSTM, incorporating the past and current temporal frames. The BLSTM outperforms the LSTM, due to the use of forward-LSTM and backward-LSTM in every BLSTM layer. The forward-LSTM is the same as the standard LSTM, which captures the interdependency between the past and current temporal frames. However, the backward-LSTM is fed by reverse input sequence, and thus the interdependency between current and future temporal frames is also utilized to achieve further improvement over the LSTM.

The CRN obtains, on average, Δ SDR = 8.81 dB, STOI = 79.49% and PESQ = 2.39, which provides more significant improvements over the DNN, S-DNN and LSTM methods. Since the CRN captures local spatial patterns of the input magnitude spectrum, it can leverage the T-F structure of the magnitude spectrum. Moreover, the LSTM layers inside the CRN exploit the temporal dependency by using past and current temporal frames. In addition, we perform experiments for the non-casual version of CRN, namely CRBN, where the BLSTM layers replace LSTM layers. The experimental results show that the CRBN offers slight improvements over the CRN method, which confirms that the interdependency between the current and future frames improves predictions by the model. The GRN outperforms the CRN by using the dilated convolutional layers.

The proposed MCGN gets the highest improvements over the baseline methods, and it achieves, on average, Δ SDR = 10.88 dB, STOI = 82.42% and PESQ = 2.58, which are almost 1.7 dB, 2.53% and 0.16 higher than those achieved by the CRN

TABLE I
SPEECH ENHANCEMENT PERFORMANCE COMPARISONS IN TERMS OF STOI AND PESQ OVER THREE DIFFERENT TYPES OF SEEN NOISES WITH DIFFERENT BASELINE METHODS AND SNR LEVELS. EACH RESULT IS THE AVERAGE VALUE OF 100 EXPERIMENTS. *Italic* TEXT REFERS TO THE PROPOSED METHODS. **BOLD** NUMBER INDICATES THE BEST PERFORMANCE

Measure	STOI (%)												
	Babble				Leopard				F16				
Noises	SNR	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture		53.86	63.05	71.66	62.86	71.68	75.57	78.92	75.39	54.39	64.07	73.37	63.94
DNN [11]		66.36	72.91	79.26	72.83	77.11	80.55	83.26	80.30	69.17	76.24	81.35	78.79
S-DNN [10]		66.80	73.66	79.63	73.36	78.34	81.54	83.98	81.29	69.77	76.46	81.86	75.97
LSTM [16]		68.78	75.81	81.54	75.38	80.76	83.32	85.40	83.16	72.13	77.86	83.39	77.79
BLSTM [15]		69.30	76.63	82.04	75.99	81.10	83.85	85.58	83.59	72.19	78.15	83.61	77.98
MRCAE [20]		65.92	72.83	78.85	72.53	77.50	80.56	83.09	81.25	69.10	75.51	80.88	75.16
CRN [22]		70.10	76.95	81.88	76.31	81.20	84.02	85.80	83.67	72.65	78.98	83.90	78.51
CRBN		70.30	77.08	81.96	76.45	81.20	84.20	85.90	83.77	73.49	79.12	84.14	78.92
GRN [15]		71.60	77.08	82.21	76.94	82.20	84.15	86.08	84.14	72.94	79.27	83.56	78.59
<i>MCGN</i>		75.02	84.43	79.99	84.10	85.78	87.31	86.55	77.49	81.56	85.60	81.55	
Measure	PESQ												
	Babble				Leopard				F16				
Noises	SNR	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture		1.28	1.52	1.81	1.53	1.75	1.99	2.22	1.97	1.31	1.54	1.83	1.56
DNN [11]		1.58	1.90	2.20	1.89	2.03	2.31	2.50	2.28	1.73	2.08	2.31	2.04
S-DNN [10]		1.69	2.00	2.28	1.99	2.25	2.45	2.67	2.46	1.82	2.11	2.35	2.09
LSTM [16]		1.82	2.15	2.44	2.14	2.41	2.61	2.80	2.61	1.97	2.28	2.52	2.25
BLSTM [15]		1.84	2.19	2.47	2.16	2.44	2.67	2.84	2.65	2.02	2.30	2.54	2.29
MRCAE [20]		1.72	2.04	2.31	2.02	2.25	2.47	2.69	2.47	1.85	2.15	2.39	2.13
CRN [22]		1.91	2.22	2.49	2.21	2.50	2.70	2.90	2.69	2.02	2.30	2.54	2.29
CRBN		1.93	2.23	2.50	2.22	2.51	2.72	2.90	2.71	2.09	2.38	2.58	2.35
GRN [15]		1.94	2.24	2.49	2.22	2.53	2.73	2.92	2.73	2.05	2.34	2.58	2.32
<i>MCGN</i>		2.16	2.43	2.65	2.41	2.70	2.88	3.04	2.87	2.23	2.47	2.70	2.47

TABLE II
p-VALUE OF THE T-TEST AT 5% SIGNIFICANCE LEVEL, BETWEEN THE PROPOSED METHOD AND THE BASELINE METHODS. H_0 DENOTES THE NULL HYPOTHESIS, AND (+) INDICATES THAT THE DIFFERENCE AMONG THE PAIR IS STATISTICALLY SIGNIFICANT AT THE 95% CONFIDENCE LEVEL

Measures	STOI		PESQ	
	<i>p</i> -value	H_0	<i>p</i> -value	H_0
Noisy	1.49E-05	(+)	4.14E-12	(+)
DNN [11]	5.22E-06	(+)	1.76E-07	(+)
S-DNN [10]	1.08E-05	(+)	3.20E-09	(+)
LSTM [16]	8.16E-05	(+)	3.49E-07	(+)
BLSTM [15]	1.93E-04	(+)	1.46E-06	(+)
MRCAE [20]	3.19E-06	(+)	1.87E-06	(+)
CRN [22]	1.44E-04	(+)	2.74E-07	(+)
CRBN	7.71E-05	(+)	1.13E-05	(+)
GRN [15]	1.01E-04	(+)	1.08E-06	(+)

method. The MCGN encodes the input magnitude spectrum in different scales. The local interdependency is captured by the convolutional sub-layers with small kernel sizes. The convolutional sub-layers with large kernel sizes are used to find the interdependency from the larger regions. By using the small and large size kernels, the receptive field of MCGN is enlarged, and the different scaled features are assigned with different weights. Furthermore, the BGRU layers are introduced to connect the multi-scale encoder and multi-scale decoder, which are capable of exploiting the interdependency of the past, current and future temporal frames. Besides, the raw data is fed to the output layer of the MCGN to learn the residual mapping relation.

We also perform the t-test between the proposed MCGN method and baseline methods, noisy mixtures for the unseen speakers with seen noises cases. The t-test results are shown in Table II. The *p*-values are all smaller than 0.05 and all the null hypothesis is (+), which indicates that the proposed MCGN method yields a statistically significant improvement over the baseline methods.

D. Unseen Speaker With Unseen Noises

Fig. 4 and Table III provide experimental results in terms of Δ SDR, STOI and PESQ for the proposed MCGN and baseline methods with unseen noises. The testing speakers are unseen in training data. The unseen testing noises are N56, N72 and White noises.

The DNN method offers slight improvements over the noisy mixture. The MRCAE outperforms the DNN method in terms of Δ SDR and PESQ, but its STOI performance is worse than that of DNN and S-DNN. These results show that the shallow structure and small channel numbers can limit the performance of MRCAE. Besides, the large size filters increase computational cost. The skip connection in S-DNN boosts enhancement performance compared to the DNN methods. The LSTM obtains further improvement by incorporating the past and current temporal information. The utilization of past, current and future temporal information in BLSTM shows advantages over the LSTM and DNN based method. The CRN method incorporates the convolutional encoder-decoder with the LSTM. The convolutional encoder-decoder takes advantage of the convolutional layer and batch normalization to provide a high-level representation of the input feature, which improves the enhancement performance. Incorporating of the BLSTM layers, the CRBN offers higher improvements over the CRN method in terms of Δ SDR, STOI and PESQ. The GRN method uses gated linear units to control the information flow, and dilated convolutional layers to expand the receptive fields. These strategies enable the GRN method to outperform the methods above.

The proposed MCGN method provides the highest improvements over all the baseline methods in terms of Δ SDR, STOI and PESQ. The t-test results in Table IV also show that the improvement of the proposed MCGN methods is statistically significant.

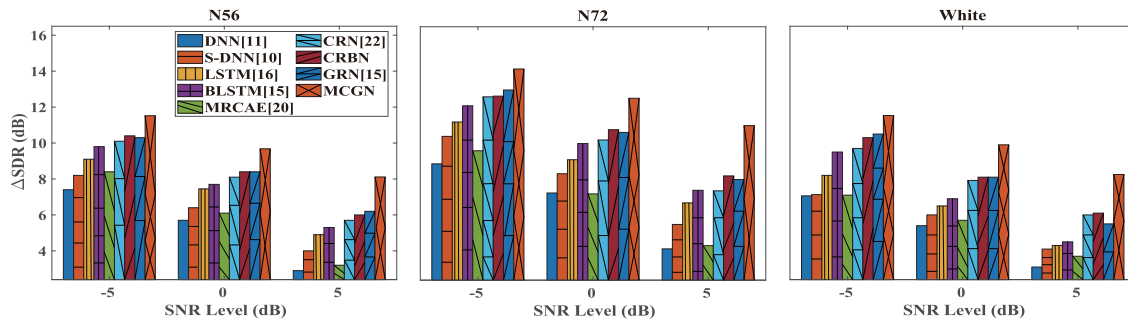


Fig. 4. Speech enhancement performance comparison in terms of Δ SDR for three unseen noises with different methods and SNR levels. Each result is the averaged value of 100 experiments.

TABLE III
SPEECH ENHANCEMENT PERFORMANCE COMPARISONS IN TERMS OF STOI AND PESQ OVER THREE TYPES OF UNSEEN NOISES WITH BASELINE METHODS AND SNR LEVELS. EACH RESULT IS THE AVERAGED VALUE OF 100 EXPERIMENTS. *Italic* TEXT REFERS TO THE PROPOSED METHODS. **BOLD** NUMBER INDICATES THE BEST PERFORMANCE

Measure Noises	STOI (%)											
	N56				N72				White			
SNR	-5dB	0dB	5dB	Avg	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	57.07	68.26	78.37	67.90	70.87	76.77	81.63	76.42	53.40	62.57	72.32	62.76
DNN [11]	72.59	78.74	84.07	78.46	75.34	81.43	85.00	80.59	62.73	70.42	77.32	70.16
S-DNN [10]	72.79	78.74	84.31	78.85	76.87	82.18	85.67	81.57	63.00	70.63	77.87	70.50
LSTM [16]	76.99	79.45	86.47	80.97	76.83	82.95	86.80	82.19	67.71	75.91	81.53	75.05
BLSTM [15]	77.64	82.49	86.87	82.33	78.05	83.50	87.12	82.89	72.93	76.43	83.91	77.76
MRCAE [20]	72.74	78.95	83.78	78.49	75.47	80.57	84.72	80.25	65.12	71.36	76.31	70.93
CRN [22]	77.88	83.37	87.09	82.78	78.55	84.12	87.24	83.30	72.90	78.93	84.34	78.72
CRBN	78.95	83.85	87.30	83.37	79.26	84.37	87.64	83.76	76.24	81.30	85.16	80.92
GRN [15]	78.19	83.60	87.33	83.04	78.96	84.27	87.60	83.61	76.31	80.80	84.59	80.57
<i>MCGN</i>	82.83	86.85	89.82	86.50	81.07	85.52	88.36	84.98	78.26	83.81	87.75	83.27
Measure Noises	PESQ											
	N56				N72				White			
SNR	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture	1.14	1.31	1.57	1.34	1.58	1.83	2.06	1.83	1.04	1.21	1.47	1.24
DNN [11]	1.69	1.98	2.16	1.95	1.77	2.05	2.23	2.02	1.32	1.61	1.87	1.60
S-DNN [10]	1.74	2.05	2.20	1.98	1.86	2.11	2.32	2.10	1.34	1.69	1.94	1.66
LSTM [16]	1.93	2.17	2.36	2.16	1.87	2.14	2.40	2.14	1.59	1.96	2.26	1.94
BLSTM [15]	1.97	2.20	2.39	2.18	1.92	2.20	2.45	2.19	1.81	2.19	2.45	2.15
MRCAE [20]	1.83	2.06	2.22	2.04	1.93	2.16	2.35	2.15	1.60	1.85	2.09	1.85
CRN [22]	1.92	2.22	2.49	2.21	1.98	2.20	2.41	2.20	1.90	2.21	2.48	2.20
CRBN	2.05	2.27	2.44	2.25	1.99	2.25	2.47	2.24	2.04	2.28	2.54	2.29
GRN [15]	2.01	2.24	2.43	2.23	2.01	2.25	2.53	2.26	2.14	2.35	2.56	2.35
<i>MCGN</i>	2.22	2.40	2.58	2.40	2.14	2.40	2.63	2.39	2.24	2.55	2.84	2.54

TABLE IV
p-VALUE OF THE T-TEST AT 5% SIGNIFICANCE LEVEL, COMPARISON OF PROPOSED METHOD WITH THE BASELINE METHODS. H_0 DENOTES THE NULL HYPOTHESIS, AND (+) INDICATES THE IMPROVEMENT OF TWO PAIRS IS STATISTICALLY SIGNIFICANT AT THE 95% CONFIDENCE LEVEL

Measures	STOI		PESQ	
	<i>p</i> -value	H_0	<i>p</i> -value	H_0
Noisy	1.49E-04	(+)	2.06E-05	(+)
DNN [11]	2.97E-04	(+)	1.73E-05	(+)
S-DNN [10]	6.75E-04	(+)	3.80E-04	(+)
LSTM [16]	4.18E-04	(+)	3.05E-04	(+)
BLSTM [15]	2.24E-04	(+)	4.17E-05	(+)
MRCAE [20]	9.94E-04	(+)	2.42E-04	(+)
CRN [22]	1.89E-04	(+)	2.62E-05	(+)
CRBN	1.06E-04	(+)	1.18E-05	(+)
GRN [15]	1.94E-04	(+)	2.82E-05	(+)

The proposed MCGN method provides the highest improvements over all the baseline methods in terms of Δ SDR, STOI and PESQ. The t-test results in Table IV also show that the improvement of the proposed MCGN methods is statistically significant.

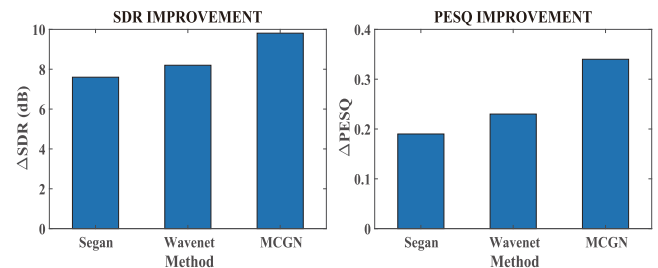


Fig. 5. Speech enhancement comparison in terms of Δ SDR and Δ PESQ for SEGAN [23], Wavenet [21] and the proposed MCGN. The enhancement results are the averaged value of 824 noisy mixtures.

E. Experiments on Published Dataset

We also evaluate the proposed MCGN method on the second dataset that we mentioned earlier, i.e. the published dataset generated by the VCTK corpus. Fig. 5 shows experimental results. Note that the model size (i.e. the number of parameters) of SEGAN, Wavenet and the proposed MCGN is 193 M,

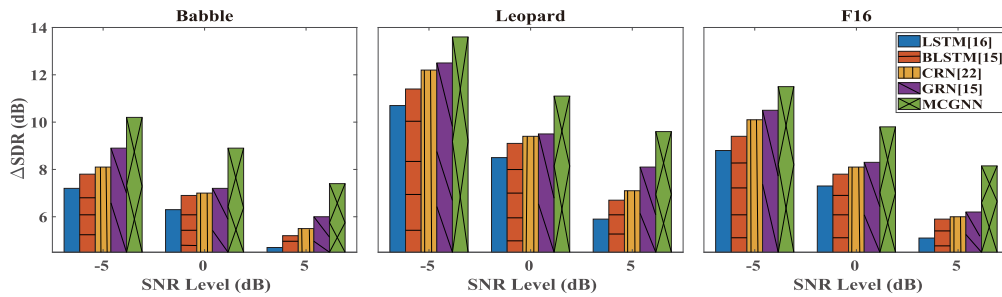


Fig. 6. Speech enhancement performance comparison in terms of Δ SDR for three types of noise with different methods and SNR levels. Each result is the averaged value of 500 experiments.

TABLE V

SPEECH ENHANCEMENT PERFORMANCE COMPARISONS IN TERMS OF STOI AND PESQ OVER THREE DIFFERENT TYPES OF SEEN NOISES WITH DIFFERENT BASELINE METHODS AND SNR LEVELS. EACH RESULT IS THE AVERAGED VALUE OF 500 EXPERIMENTS. *Italic* TEXT REFERS TO THE PROPOSED METHOD. **BOLD** NUMBER INDICATES THE BEST PERFORMANCE

Measure	STOI (%)												
	Babble				Leopard				F16				
Noises	SNR	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Noisy Mixture		56.63	66.36	75.28	66.09	73.22	78.06	82.26	77.85	57.4	66.83	75.61	66.61
LSTM [16]		70.30	78.18	83.96	77.48	81.15	84.76	87.45	84.45	74.21	80.91	85.23	80.12
BLSTM [15]		71.24	78.86	84.53	78.21	82.02	85.39	87.99	85.13	74.98	81.28	85.93	80.73
CRN [22]		71.34	79.18	84.67	78.40	82.30	85.77	88.17	85.41	75.06	81.98	86.13	81.06
GRN [15]		73.53	80.06	84.29	79.27	82.77	85.9	88.21	85.63	77.38	82.66	86.11	82.05
<i>MCGN</i>		78.46	83.19	86.97	82.87	85.01	87.45	89.45	87.30	79.56	84.27	87.64	83.82
Measure	PESQ												
Noises	SNR	Babble				Leopard				F16			
Noisy Mixture		1.38	1.65	1.96	1.66	1.77	2.09	2.37	2.08	1.31	1.55	1.83	1.56
LSTM [16]		1.88	2.25	2.56	2.23	2.44	2.69	2.92	2.68	2.02	2.37	2.63	2.34
BLSTM [15]		1.91	2.29	2.60	2.26	2.50	2.75	2.98	2.74	2.10	2.41	2.67	2.39
CRN [22]		1.93	2.31	2.61	2.28	2.53	2.79	3.00	2.77	2.14	2.45	2.69	2.43
GRN [15]		2.03	2.33	2.62	2.33	2.53	2.80	3.01	2.78	2.20	2.47	2.70	2.46
<i>MCGN</i>		2.27	2.55	2.77	2.53	2.71	2.93	3.10	2.91	2.35	2.60	2.81	2.59

34.3 M, 77.5 M respectively. The no-casual, dilated convolutions controlled by the Sigmoid gate in every layer help to enlarge the receptive fields of every kernel, and thus to utilize the interdependency among input features. The future samples help the Wavenet to perform better. Our MCGN method produces substantially better enhancement performance, since the MCFR model provides weighted multi-scale feature in every layer, and captures the interdependency among different frames including future frames.

F. Additional Experiments

Figs. 6 & 7 and Tables V & VI provide experimental results in terms of Δ SDR, STOI and PESQ for the proposed MCGN and four baseline methods (i.e. LSTM, BLSTM, CRN and GRN) with seen and unseen noises, for the larger dataset (i.e. 50 000 training signals and 2500 testing signals for each SNR level, described in Section III.A).

It can be observed that the proposed MCGN method performs better than all the baseline methods, and shows similar trends as for the smaller dataset tested earlier. All the methods provide some improvements over the noisy mixtures, which indicate that they are effective for speech enhancement with seen and unseen noises. The BLSTM provides more improvements than LSTM, since it uses additional information from the future frames, in contrast to the information from only current and previous frames used in LSTM. The CRN uses the CED to

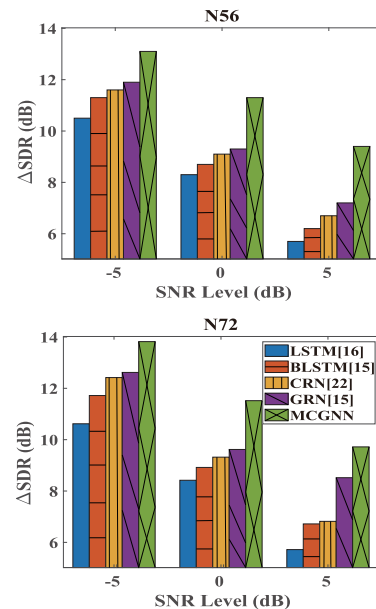


Fig. 7. Speech enhancement performance comparison in terms of Δ SDR for two unseen noises with different methods and SNR levels. Each result is the averaged value of 500 experiments.

capture local T-F patterns from input noisy mixtures, also uses the LSTM layers to relate the past frames with current frames, thus offering higher improvements than the LSTM and BLSTM.

TABLE VI
SPEECH ENHANCEMENT PERFORMANCE COMPARISONS IN TERMS OF STOI AND PESQ OVER TWO DIFFERENT TYPES OF UNSEEN NOISES WITH BASELINE METHODS AND SNR LEVELS. EACH RESULT IS THE AVERAGED VALUE OF 500 EXPERIMENTS. *Italic* TEXT REFERS TO THE PROPOSED METHOD. **BOLD** NUMBER INDICATES THE BEST PERFORMANCE

Measure	STOI (%)							
	N56				N72			
Noises	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Mixture	57.83	67.19	76.02	67.01	79.19	78.37	83.44	80.33
LSTM [16]	77.29	82.25	85.93	81.82	81.75	85.62	88.36	85.24
BLSTM [15]	77.59	82.54	86.45	82.19	82.59	86.18	89.11	85.96
CRN [22]	77.90	83.62	87.24	82.92	82.61	86.81	89.25	86.22
GRN [15]	78.50	83.88	87.56	83.31	83.00	87.4	89.35	86.58
<i>MCGN</i>	83.87	87.20	89.67	86.91	85.57	88.55	90.54	88.22
Measure	PESQ							
	N56				N72			
Noises	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
Mixture	1.20	1.33	1.56	1.36	1.57	1.84	2.10	1.84
LSTM [16]	2.03	2.22	2.40	2.22	2.15	2.39	2.58	2.37
BLSTM [15]	2.05	2.26	2.44	2.25	2.23	2.42	2.62	2.42
CRN [22]	2.07	2.31	2.47	2.28	2.28	2.46	2.65	2.46
GRN [15]	2.08	2.32	2.45	2.28	2.29	2.46	2.67	2.47
<i>MCGN</i>	2.27	2.51	2.64	2.47	2.42	2.61	2.82	2.62

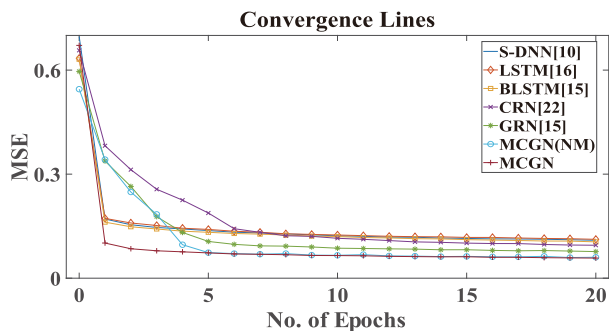


Fig. 8. Mean squared errors over training epochs for S-DNN, LSTM, BLSTM, CRN, GRN, MCGN, and MCGN(NM) on the testing set. The MCGN(NM) represents the delete the multi-scale output layers, only use the normal output layer. All models are evaluated with a testing set of unseen speakers.

The GRN shows advantage over LSTM, BLTM and CRN, due to the employment of the dilated 2-D convolutional layers for expanding the receptive fields in the T-F domain, and the gated convolution to control the information flow between layers.

The proposed MCGN method employs multi-scale 2D-convolutional layers to enlarge the receptive fields in the T-F domain, as a result, the features extracted are in different scales, capturing both local and contextual information. The multi-scale feature is assigned with different weights to provide a better feature representation. Furthermore, the BGRU layers are utilized to model the interdependency among the past, current and future frames. In summary, these results further confirm that the MCGN outperforms baseline methods.

G. Convergence Lines and Spectrums

Fig. 8 demonstrates the testing MSEs of the baseline methods and the proposed MCGN and MCGN without multi-scale output (MCGN(NM)) layers over epochs. It can be seen that the MCGN converges faster than the baseline methods and reaches the lowest MSE. After 20 epochs training, the MCGN and MCGN(NM) offer similar MSEs, but the convergence speed of MCGN is faster than MCGN(NM) at 1-5 epochs. This suggests that the

TABLE VII
COMPONENT ANALYSIS

Measures	Δ SDR	STOI	PESQ	Parameters
Full	10.20	81.40	2.40	77.5
No Bottleneck	10.39	81.60	2.43	133.4
No FC	10.24	81.51	2.40	123.7
No CL	9.27	77.25	2.23	41.9
No MCFR	9.12	77.42	2.20	27.9
No FR	9.61	79.87	2.32	68.8

multi-scale feature representation may also help improve the convergence speed of the algorithm, apart from improving its enhancement performance.

We plot a set of spectrums in Fig. 9. It can be seen that the baseline methods and the proposed MCGN method provide different enhancement performance in terms of reconstruction of target speech. The spectrums of the proposed MCGN method are closer to the spectrums of the target speech, which again confirms that the MCGN outperforms the baseline methods.

H. Component Analysis

We also conduct a series of experiments to investigate the efficiency of different components in the proposed model. In the component analysis, the ablation experiments are performed by removing different components to show how it affects the enhancement performance.

Table VII provides the experimental results of using various components in terms of the Δ SDR, STOI, PESQ and the parameters (million). Full means the full MCGN framework. No bottleneck represents removing the bottleneck layers in MCGN. No FC represents removing the fully connected layers in MCGN. No MCFR means using the single kernel in each encoder-decoder layer. No CL represents removing the connection layers that include a dense layer and two BGRU layers. No FR denotes removing feature recalibration, which means that the different scaled features use the same weight and are concatenated directly.

The bottleneck layers employ fewer channels than previous layers to compress the information from previous convolutional layers, and this can reduce the computational cost with slight information loss, as shown in the experimental results. Unlike bottleneck layers in the convolutional encoder and decoder, the FC layer with non-linear activation can produce a compact representation of the encoder output before the BGRU layer is applied. The bottleneck and FC layers help capture global information from the mixture. In addition, the interdependency among the past, current and future frames is captured by the BGRU layers. Therefore, the CL can employ BGRU and FC layers to provide improvements of enhancement performance and parameter efficiency. The results also show that the MCFR module can improve the performance by capturing the features in different scales using paralleled kernels of different size.

Fig. 10 shows the weights obtained by feature recalibration in the last layer of the multi-scale decovolutional layer. The color-bar shows the weight values, and the deeper color represents a smaller value. Comparing Fig. 10 with Fig. 9 (A), (B), we can observe that the weights of high values capture the target

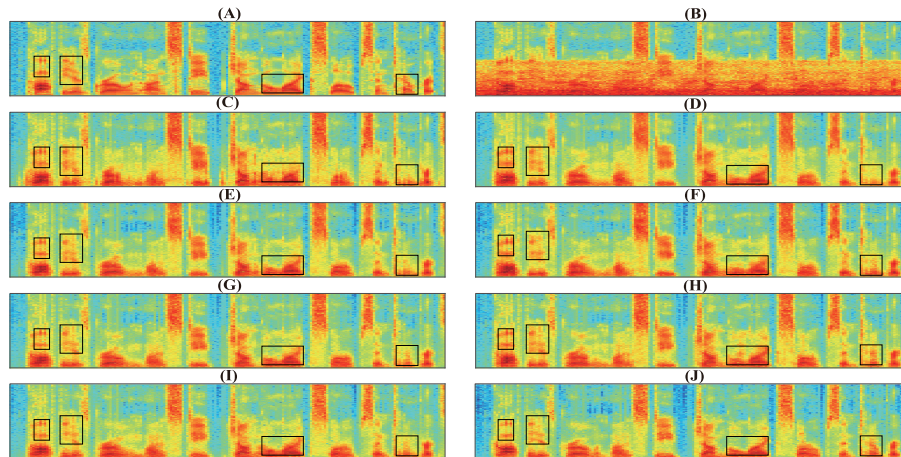


Fig. 9. Spectrums of different signals: (A) clean speech; (B) noisy speech mixture; (C) enhanced speech by S-DNN [10]; (D) enhanced speech by LSTM [16]; (E) enhanced speech by BLSTM [15]; (F) enhanced speech by the proposed MRCAE [20] (G) enhanced speech by the proposed CRN [22]; (H) enhanced speech by the proposed CRBN; and (I) enhanced speech by the proposed GRN [15]. (J) enhanced speech by the proposed MCGN.

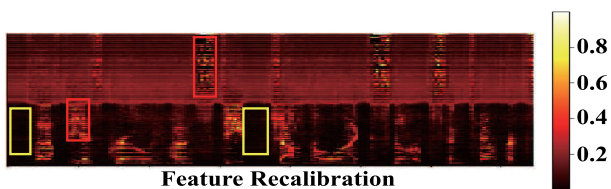


Fig. 10. Weights obtained by feature recalibration are shown as a hot-map, where the horizontal and vertical axis denote the time and frequency, respectively, and the color-bar shows the values of the weights.

TABLE VIII
KERNEL SIZE ANALYSIS

Filter Size	Δ SDR	STOI	PESQ
1×2	10.55	72.07	1.71
2×2	10.72	72.21	1.72
2×3	10.76	72.30	1.73
4×5	10.88	72.37	1.73
5×5	11.16	72.97	1.76
7×7	11.18	73.15	1.77
11×11	11.23	73.07	1.75
Multi-Kernel	11.72	76.21	1.92

speech very well. For example, the areas highlighted with the red blocks represent speech components, while those highlighted with yellow blocks represent components from noise. It can be observed that the feature recalibration tends to assign the features from speech with higher weights, and features from noise with lower weights. Therefore, the feature recalibration helps suppress noise and improve reconstruction of the target speech.

I. Kernel Size Analysis

We perform further experiments to analyse the relation between enhancement performance and kernel sizes with unseen noises. These experiments use kernel size varied from 1×2 to 11×11 , thus exploiting different receptive fields in the T-F domain. Table VIII provides the experimental results in terms

of Δ SDR, STOI, and PESQ. The enhancement performances increase with the increase in the kernel size, e.g. from 1×2 to 7×7 , but then starts to saturate for the further increase to 11×11 . However, the performance difference is relatively small.

A larger kernel size, such as 7×7 , can provide a larger receptive field, which generates the T-F feature map from a larger region i.e. contextual information, which may be effective in mitigating noise, and a smaller kernel size such as 1×2 captures the feature map among a smaller region i.e. local information, thus effective in retaining the detailed T-F structure. This appears to be consistent with the analysis in [15], [22]. Unlike the BGRU layers which capture the interdependency among time frames (i.e. time-domain), the 2D-convolutional layers allow the expansion along both time and frequency.

As shown in Table VIII, the performance is also dependent on the choice of the kernel size. When the kernel size is larger than 7×7 , performance may decrease in terms of STOI and PESQ. Using paralleled multi-kernel helps the model to capture the features in different scales, thus exploiting both local and contextual information, and to offer better enhancement performance with unseen noises, as in our proposed method.

To interpret the use of different kernel sizes, we have provided an example of the feature map obtained by using kernels of different sizes in the first multi-scale convolutional layer, as shown in Fig. 11, using kernels of size 1×2 , 3×4 , and 7×7 . It can be seen, although the kernel at each scale extracted both speech and noise components, as shown in the regions highlighted with blue and black, the feature maps obtained with these kernels characterise different receptive fields, for example, with the large kernel, more heavy smoothing is applied which is effective in mitigating the impact of noise, while the use of a small kernel can retain the fine structure of the spectrum. Therefore, using a bank of kernels, the system has a better chance to capture and distinguish the features from speech and noise, thus further improves the speech enhancement performance.

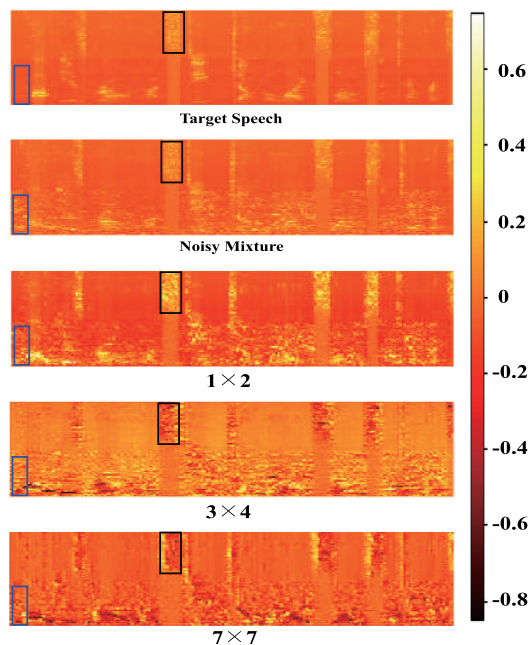


Fig. 11. Feature maps in MCGN with different kernel sizes. Also, the magnitude spectra of target speech and noisy mixture are provided. The horizontal and vertical axes denote the time and frequency, respectively. The color-bar shows the L2-normalized values. Example of components of speech and noise are highlighted with black and blue blocks, respectively.

IV. CONCLUSION

We have presented a new framework for single channel speech enhancement. The proposed MCGN introduced several novel strategies to improve the enhancement performance and computational efficiency of the neural network algorithm. Firstly, we introduced the MCFR structure to extract the features in different scales, capturing both the local and contextual information from the speech mixtures. In addition, the feature recalibration network was implemented using a gating function to control the information flow, which assigns higher weights to speech components and lower weights to noise components, and thus improves the reconstruction of target speech by suppressing the noise from noisy mixtures. Secondly, we introduced the bottleneck convolutional and deconvolutional layers to reduce information flow dimension in encoder and decoder, but to retain the information. Thirdly, the efficiency connection module was introduced. The fully connected layer was used to reduce the dimension of the output of the convolutional encoder. The BGRU was exploited to capture the interdependency among the past, current and future temporal frames, which provides comparable performance with fewer parameters than BLSTM. Finally, we introduced the multi-scale convolutional output layer, then summed the multi-scale outputs to accelerate the convergence speed. A variety of noises were used to examine the enhancement performance of the system. The unseen speakers with the seen and unseen noises were exploited to evaluate the efficacy of the proposed method. The experimental results confirmed the improved performance of the proposed method over the state-of-the-art baseline methods.

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor and the anonymous reviewers for their valuable input to improving the paper.

REFERENCES

- [1] D. L. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectr.*, vol. 54, no. 3, pp. 32–37, Mar. 2017.
- [2] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.
- [3] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE J. Sel. Top. Signal Process.*, vol. 4, no. 5, pp. 895–910, Oct. 2010.
- [4] Y. Sun, Y. Xian, W. W. Wang, and S. M. Naqvi, "Monaural source separation in complex domain with long short-term memory neural network," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 2, pp. 359–369, May 2019.
- [5] B. Rivet, W. W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 125–134, May 2014.
- [6] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 3, pp. 197–210, Jun. 1978.
- [7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
- [8] Y. X. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.
- [9] K. Han, Y. X. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 189–198, Jun. 2015.
- [10] M. Tu and X. X. Zhang, "Speech enhancement based on deep neural networks with skip connections," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5565–5569.
- [11] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.
- [12] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2002, pp. I-529–I-532.
- [13] 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.
- [14] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural network for robust speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7092–7096.
- [15] K. Tan 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.
- [16] J. T. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *J. Acoustical Soc. America*, vol. 141, no. 6, pp. 4705–4714, 2017.
- [17] S. Hochreiter and J. Schmidhuber, *Long Short-Term Memory*. Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
- [18] F. Weninger *et al.*, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust ast," in *Proc. Int. Conf. Latent Variable Anal. Signal Separation (LVA/ICA)*, 2015, pp. 91–99.
- [19] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 1993–1997.
- [20] E. M. Grais, D. Ward, and M. D. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," in *Proc. Eur. Signal Process. Conf.*, 2018, pp. 1591–1595.
- [21] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5069–5073.
- [22] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3229–3233.

- [23] P. Santiago, B. Antonio, and S. Joan, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Int. Speech Commun. Assoc.*, 2017, pp. 3642–3646.
- [24] J. Y. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. Conf. Neural Inf. Process. Syst.*, 2014, [Online]. Available: <https://arxiv.org/abs/1412.3555>
- [25] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
- [26] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.
- [27] A. L. Mass, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn. Workshop Deep Learn. Audio, Speech and Lang. Process.*, 2013, pp. 1–6.
- [28] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [29] C. Szegedy, V. Vanhoucke, S. Ioffe, and Z. W. J. Shlens, "Rethinking the inception architecture for computer vision," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.
- [30] K. Y. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.
- [31] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2342–2350.
- [32] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "Darpa timit acoustic-phonetic continuous speech corpus," *Nat. Inst. Standards Technol.*, Tech. Rep. NISTIR 4930, Gaithersburg, MD, USA, vol. 93, 1993.
- [33] G. N. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.
- [34] A. Varga and H. 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.
- [35] C. Veaux, J. Yamagishi, and K. MacDonald, "University of Edinburgh. The Centre for Speech Technology Research (CSTR)," 2016. [Online]. Available: <https://doi.org/10.7488/ds/1994>
- [36] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. 21st Int. Congr. Acoust.*, Montreal, Canada, Acoustical Soc. America, 2013. [Online]. Available: https://hal.inria.fr/hal00796707/file/thiemann_demand.pdf
- [37] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
- [38] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
- [39] 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.
- [40] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015. [Online]. Available: <https://arxiv.org/abs/1412.6980>



Yang Xian (Student Member, IEEE) received the B.Sc. degree in 2014 from the Zhengzhou University, Zhengzhou, China, and the M.Sc degree in 2016 from Newcastle University, Newcastle, U.K. He is currently working toward the Ph.D. degree with Intelligent Sensing and Communications Research Group, School of Engineering, Newcastle University. His research interests include audio signal processing, speech source separation, and enhancement based on deep learning.



Yang Sun (Member, IEEE) received the Ph.D. degree with Intelligent Sensing and Communications Research Group, School of Engineering, Newcastle University, Newcastle, U.K., in 2019. He is currently a Postdoctoral Researcher with Big Data Institute, University of Oxford, Oxford, U.K., developing methods for brain lesion segmentation with multiple sclerosis. His research interests include audio signal processing and biomedical image processing based on deep learning.



Wenwu Wang (Senior Member, IEEE) received the B.Sc., M.E., and Ph.D. degrees from the College of Automation, Harbin Engineering University, Harbin, China, in 1997, 2000, and 2002, respectively. From 2002 to 2003, he was with King's College London, London, U.K., from 2004 to 2005, with Cardiff University, Cardiff, U.K., from 2005 to 2006, with Tao Group Ltd. (now Antix Labs Ltd.), from 2006 to 2007, with Creative Labs. In May 2007, he joined the University of Surrey, where he is a Professor in Signal Processing and Machine Learning, and a

Co-Director of the Machine Audition Lab, Centre for Vision Speech and Signal Processing. He is also a Guest Professor with the Qingdao University of Science and Technology, Qingdao, China. In 2008, he was a Visiting Scholar with Ohio State University, Columbus, OH, USA. He has coauthored more than 250 publications in his research interests, which include blind signal processing, sparse signal processing, audio-visual signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. Prof. Wang is a Senior Area Editor and was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2014 to 2018. He is an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING, and the EURASIP *Journal on Audio Speech and Music Processing*. He was a Publication Co-Chair of the ICASSP 2019, Brighton, U.K. He is a member of the IEEE Signal Processing Theory and Methods Technical Committee and a Member of the IEEE Machine Learning for Signal Processing Technical Committee.



Syed Mohsen Naqvi (Senior Member, IEEE) received the Ph.D. degree in signal processing from Loughborough University, Loughborough, U.K., in 2009, and his Ph.D. thesis was on the EPSRC U.K.-funded project. He was a Postdoctoral Research Associate with the EPSRC U.K.-funded projects and a Research Excellence Framework (REF) Lecturer from 2009 to 2015. Prior to his postgraduate studies in Cardiff and Loughborough Universities, he was with the National Engineering and Scientific Commission (NESCOM) of Pakistan from 2002 to 2005. He is

currently an Associate Professor/Senior Lecturer in Signal and Information Processing with the School of Engineering, Newcastle University, Newcastle, U.K. He is leading Intelligent Sensing Laboratory, Newcastle University, with major research focused on multimodal processing for human behavior analysis, multitarget tracking, and source separation all for machine learning. He organized special sessions in FUSION, delivered seminars, and was a Speaker with University Defence Research Collaboration Summer Schools from 2015 to 2017. He has authored or coauthored 150 publications with the main focus of his research on audio-visual signal and information processing, machine learning and perception, reliable artificial intelligence, and action recognition and anomaly detection. He is an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and Elsevier *Journal on Signal Processing*. He is a fellow of the Higher Education Academy.