Multi-Resolution Convolutional Residual Neural Networks for Monaural Speech Dereverberation

Lei Zhao, Wenbo Zhu, Shengqiang Li, Hong Luo, Xiao-Lei Zhang, *Senior Member, IEEE*, and Susanto Rahardja, *Fellow, IEEE*

***Abstract*—It is known that the reverberant speech in different acoustic environments varies according to reverberation time. However, most deep learning based speech dereverberation methods rely on a single deep model to learn the context information. It may make the deep model biased to only part of the reverberant time durations. In this paper, we propose a multi-resolution framework to address this issue. The framework integrates the dereverberant ability of multiple deep subnetworks with different time resolutions into a unified model by transferring the dereverberant information from high-resolution subnetworks to low-resolution subnetworks. By doing so, the unified model can perform well in both long and short reverberant time. We further propose two implementations of the framework based on advanced convolutional residual neural networks. The first implementation, named multi-resolution UNet, uses our new implementation of UNet based on convolutional blocks as the dereverberation subnetwork. The second implementation, named multi-resolution stacked convolutional blocks, uses our new stacked convolutional blocks as the subnetwork. Experimental results in both simulated and real-world environments show that the proposed algorithms outperform the state-of-the-art dereverberation methods in terms of both the evaluation metrics for speech dereverberation and word error rate (WER) for speech recognition.**

***Index Terms*—Multi-resolution framework, speech dereverberation, UNet, stacked convolutional blocks.**

I. Introduction

SPEECH is usually corrupted by reverberation from surface reflections in indoor environments [1]. Strong reverberation significantly degrades speech intelligibility and speech quality for human listeners, especially for hearing impaired people. Reverberation may also dramatically deteriorate the intelligence of machines, like automatic speech recognition (ASR). Therefore, speech dereverberation becomes an importance topic of speech processing in the last decades. This paper focuses on monaural speech dereverberation.

Conventional speech dereverberation algorithms are usually unsupervised signal processing methods. In [2], Lebart et al. proposed *spectral subtraction* where an exponential decay model is used to model reverberation. In [3], Wu and Wang proposed a two-stage algorithm, which first suppresses late reverberation by spectral subtraction, and then reduces early reverberation via an inverse filter. In [4], Yoshioka et al. proposed the weighted prediction error (WPE) dereverberation algorithm, which first obtains a set of linear prediction filters based on historical frames and then obtains dereverberant speech by subtracting the filtered speech from the reverberant speech. In [5], Kalman filters, which are built by an expectation-maximization algorithm, are applied to speech dereverberation.

With the fast development of deep learning technologies, speech dereverberation based on supervised deep learning receives much attention. It demonstrates good performance in strong reverberant scenarios. The first research respect of this direction mainly focuses on the training objectives and acoustic features. In [6], Han et al. proposed to learn *spectral mapping* from the log magnitude spectrogram of reverberant speech to the corresponding anechoic speech by a deep neural network (DNN). In [7], they further extended the above approach to the tasks of both denoising and dereverberation, which learns a mapping from the spectrogram of noisy and reverberant speech to its clean speech counterpart. In [8], Zhao et al. observed that spectral mapping is more effective for dereverberation than T-F masking, whereas masking works better than mapping for denoising. Therefore, they proposed a two stage DNN, where the first stage performs ratio masking for denoising and the second stage spectral mapping for dereverberation.

Another important research respect is the design of the network structure. Early work used feedforward DNN where the contextual information is modeled by grouping the contextual frames into a long feature [6], [7]. Later, recurrent neural networks (RNN) was used to extract long-term information to perform speech dereverberation [9]. Some related work also used long short-term memory (LSTM) to model the contextual information [10], [11], [12]. Recent state-of-the-art

Manuscript received 3 April 2023; revised 12 August 2023 and 3 March 2024; accepted 25 March 2024. Date of publication 4 April 2024; date of current version 19 April 2024. This work was supported in part by the National Science Foundation of China (NSFC) under Grant 62176211, and in part by the Project of the Science, Technology, and Innovation Commission of Shenzhen Municipality, China, under Grant JCYJ20210324143006016 and Grant JSGG20210802152546026. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yu Tsao. *(Lei Zhao and Wenbo Zhu are co-first authors.) (Corresponding authors: Xiao-Lei Zhang; Susanto Rahardja.)*

Lei Zhao, Wenbo Zhu, Shengqiang Li, and Xiao-Lei Zhang are with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Research and Development, Institute of Northwestern Polytechnical University, Shenzhen 518063, China (e-mail: zhao_lei@mail.nwpu.edu.cn; wbzhu@mail.nwpu.edu.cn; shengqiangli@mail.nwpu.edu.cn; xiaolei.zhang@nwpu.edu.cn).

Hong Luo is with the China Mobile (Hangzhou) Information Technology Company Ltd., Hangzhou 311199, China (e-mail: luohong@cmhi.chinamobile.com).

Susanto Rahardja is with the School of Marine Science and Technology,, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: susantorahardja@ieee.org).

Digital Object Identifier 10.1109/TASLP.2024.3385270

2329-9290 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.

models are built on UNet. The UNet, which was first proposed for biomedical image segmentation [13], consists of multiple upsampling layers, downsampling layers, and skip connections between them. In [14], Ernst et al. introduced UNet to speech dereverberation, where the skip connection is important to avoid the loss of the essential low level information during the downsampling process. In [15], Kothapally et al. observed that the skip connection may limit the learning ability of UNet due to the granularity mismatch between the features. To address this issue, they added some convolutional layers into the skip connection. The new skip connection is named *SkipConv block*.

The work above mainly focused on estimating the magnitude spectrogram of a noisy speech signal, leaving its noisy phase spectrogram unprocessed. To address this issue, conducting enhancement and dereverberation in the complex domain is a recent trend. First, some methods try to recover the clean phase spectrogram in the polar coordinates of the complex spectrogram. For example, in [16], Zheng et al. proposed to address the phase wrapping problem by estimating the instantaneous frequency deviation (IFD) of the clean phase spectrogram, and then reconstruct the clean phase spectrogram from IFD and a reliable initial phase estimate. Another type of methods recover the clean complex spectrogram by enhancing its real and imaginary parts respectively. Specifically, in [17], Williamson et al. performed denoising and dereverberation simultaneously in the complex domain by estimating a complex ideal ratio mask in both the real and imaginary domains. Based on [17], Zhang et al. proposed a weighted magnitude-phase loss function in [18], which outperforms the regular mean squared error for speech dereverberation. The third type of methods use complex-valued networks to predict the complex-valued spectrograms of short-time Fourier transform (STFT) directly. In [19], Choi et al. proposed DCUNet, which is a UNet-based model incorporating well-defined complex-valued building blocks to deal with complex-valued spectrograms. Inspired by [19], Hu et al. proposed a deep complex convolution recurrent network in [20], which utilizes complex LSTM module to handle complex-valued operations.

Similar to the work on the complex domain, studies on time domain, which directly take original audio waveforms as the input of the network without utilizing STFT, is another popular topic. In [21], Stoller et al. proposed Wave-U-Net which takes one-dimensional audio signals directly as its input. By utilizing a one-dimensional convolution operator, it resamples feature maps at different time scales. In [22], Luo et al. proposed Conv-TasNet, which uses a linear encoder to learn a representation that is comparable to the STFT spectrogram, and then uses an advanced network to enhance the learned representation. In [23], Wang et al. proposed a transformer neural network based on UNet. It adopts dilated-dense blocks in both the encoder and decoder layers of the UNet to strengthen the feature propagation and enlarge the receptive field. It also utilizes transformer modules to extract contextual information.

Recently, the study on contextual information and reverberation time is becoming a focus of deep learning based speech dereverberation. In [24], Wu et al. observed that selecting appropriate frame length and frame shift based on the reverberation time in terms of T_{60} can improve the performance. Based on the observation, they incorporated T_{60} into the feature selection stage. Inspired by [24], Zhao et al. [25] argued that the parameters related to reverberation time can be obtained from the reverberant speech by encoding the relationship between the input features extracted from the reverberant speech rather than using a reverberation time estimator. Therefore, they applied a self-attentive mechanism to extract the correlation between the features in different time steps, which can produce dynamic representations varying along with different reverberant environments. In [26], Wang et al. proposed an effective method to exploit contextual information for environment-aware speech dereverberation in real reverberant environments. The method is a DNN-based temporal-contextual attention approach that adaptively attends to contextual information. In addition, considering that the room impulse response decays faster at high frequencies than those lower, the authors also proposed a sub-band-based timing attention method. In [27], Zhou et al. proposed a new learning objective based on reverberation time to reduce prediction errors as well as signal distortions.

Although the above advanced context-aware learning approaches are able to address the variation of the reverberation time in different acoustic environments to some extent, how to identify the acoustic environments or estimate the reverberation time is a hard issue. To prevent this hard problem, in this paper, we propose to integrate multiple dereverberation subnetworks that serve the best for different reverberation time into a unified model by a multi-resolution processing framework. Under this framework, the unified model could make the subnetworks complement with each other for improving the performance in a wide range of reverberation time, without resorting to identifying the environments apparently. The novelty and contribution of our work are summarized as follows:

- *A multi-resolution framework was proposed for monaural speech dereverberation:* The framework contains multiple dereverberation branches. The branches first partition an input utterance into different number of non-overlapping speech segments that have an equal length. The number of segments are called *resolution*. Then, the dereverberation process is executed from the branch with the highest resolution to the branch with the lowest resolution in sequence. Each branch receives dereverberation information from its previous branch for improving its dereverberation performance, and then generates dereverberation information for its successive branch. The output of the branch with the lowest resolution is used as the final dereverberation result. With this information transfer function (ITF), the dereverberation ability of different branches is integrated without having to identify the reverberation time apparently.
- *Two advanced implementations of the framework were proposed:* Each of the implementations contains four successive blocks, which are the convolutional block (CB), ITF, dereverberation subnetwork, and mask block (MB), respectively. Particularly, ITF is responsible to receive dereverberation information from another branch, while MB is responsible to generate dereverberation information for other branches. All components are convolutional

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.

residual neural networks (CRNNs). The two implementations differ in the dereverberation subnetwork. The first one is a new UNet implementation based on CB for speech dereverberation. The other one is a stack of multiple CBs, which is, to our knowledge, also a new speech dereverberation model. The two implementations are denoted as multi-resolution UNet (MR-UNet) and multi-resolution stacked convolutional blocks (MR-SCB) respectively.

- *State-of-the-art performance on speech dereverberation was achieved:* We have compared the proposed methods with the representative WPE [28] as well as eight state-of-the-art deep learning based speech dereverberation methods [10], [11], [13], [15], [23], [29], [30], [31], in both simulated and real-world highly-reverberant environments. Experimental results show that the proposed methods outperform the comparison methods significantly in a number of evaluation metrics. Ablation studies further demonstrate the strong ability of the proposed methods in dealing with different reverberation time. Moreover, the proposed MR-SCB behaves quite similar to MR-UNet.
- *Successful applications to far-field speech recognition were made:* After applying the comparison methods to two conformer-based ASR systems. Experimental results show that the proposed MR-UNet yields a relative word error rate reduction of 37.23% over the best referenced method [15] when the ASR system was trained with clean speech, and 9.63% over the best referenced method [28] when the ASR system was trained with both clean speech and reverberant speech.

The rest of the paper is organized as follows. In Section II, we present the motivation of the proposed methods. In Section III, we describe the proposed framework as well as its two implementations. In Section IV, we introduce the experimental setup. In Sections V and VI, we present the experimental results on simulated data and real-world data respectively. In Section VII, we apply the proposed methods to speech recognition. Finally, we conclude the paper in Section VIII.

II. Motivation and Related Work

In real world applications, a common way of training a dereverberation network is to gather or generate reverberant utterances in various adverse environments with a wide range of reverberation time as the training data, which is known as the *noise-independent training* or *multi-condition training*. However, the effect of the reverberant utterances to the training loss is usually different. For example, some training utterances may contribute more to the training loss reduction than the other training utterances, which makes the trainable parameters of the model change greater than that with the other training utterances. As a result, the effectiveness of the model will bias towards similar test conditions of these training utterances. Therefore, it is needed to integrate multiple dereverberation networks that are suitable to different ranges. This problem is particularly serious in speech dereverberation, since the reverberation time varies in a wide range. Moreover, a room impulse response (RIR) function consists of the impulse responses of the direct sound, early reflections, and late reflections. The power of the early and late reflections is also fundamentally different. Making average effort to all reflections may not be effective enough as well.

Therefore, it is needed to integrate multiple dereverberation networks that are suitable to different ranges of reverberation time into a unified one. Multi-resolution processing scheme provides this opportunity. It has demonstrated the effectiveness in many related tasks, including speech separation [32], speech enhancement [33], image classification [34], image restoration [35], [36], [37], etc. Although the implementations of the scheme in the aforementioned applications may be different, their motivation and core idea are similar, which motivate our work as well.

III. Multi-Resolution Convolutional Residual Neural Networks

In this section, we first present the signal model of reverberant speech in Section III-A, and then we show the proposed multi-resolution dereverberation framework in Section III-B, and finally we propose two implementations of the framework in Sections III-C to III-G.

A. Signal Model

Given a RIR $h[n]$ where n denotes time, a reverberant speech signal received by an omni-directional microphone can be modeled as:

$$y[n] = s[n] * h[n] \tag{1}$$

where $*$ denotes the convolution operation, $s[n]$ denotes clean speech, and $y[n]$ denotes reverberant speech. Note that this paper focuses on dereverberation without further considering additive noise. Equation (1) can be further written as:

$$\begin{aligned} y[n] &= s[n] * h_d[n] + s[n] * h_r[n] \\ &= x[n] + r[n] \end{aligned} \tag{2}$$

where $x[n]$ denotes the direct sound, $r[n]$ denotes the reverberant noise of $y[n]$, and $h_d[n]$ and $h_r[n]$ represent the impulse response functions for direct sound and reverberation respectively. The objective of dereverberation is to recover the direct sound $x[n]$ from the reverberant speech $y[n]$.

We transform the speech signal in the time domain to a spectrogram in the time-frequency domain by STFT:

$$\mathbf{y}_n = \mathbf{x}_n + \mathbf{r}_n, \quad \forall n = 1, \ldots, N \tag{3}$$

where N is the number of frames of the speech signal, $\mathbf{y}_n$, $\mathbf{x}_n$, and $\mathbf{r}_n$ are the complex spectrograms of the reverberant speech, direct sound, and reverberant noise respectively.

The proposed method works on the polar coordinates of the STFT feature, though rectangular coordinates can be applied as well by e.g. [17], [38]. In the polar coordinate system, we denote $|\mathbf{y}|$ and $\angle\mathbf{y}$ as the magnitude spectrogram and phase spectrogram of $\mathbf{y}$ respectively. Similarly, we denote $|\mathbf{x}|$ and $\angle\mathbf{x}$ as the magnitude spectrogram and phase spectrogram of $\mathbf{x}$ respectively. The proposed method takes $|\mathbf{y}|$ as the input, and aims to estimate $|\mathbf{x}|$. In the prediction stage, after getting

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.

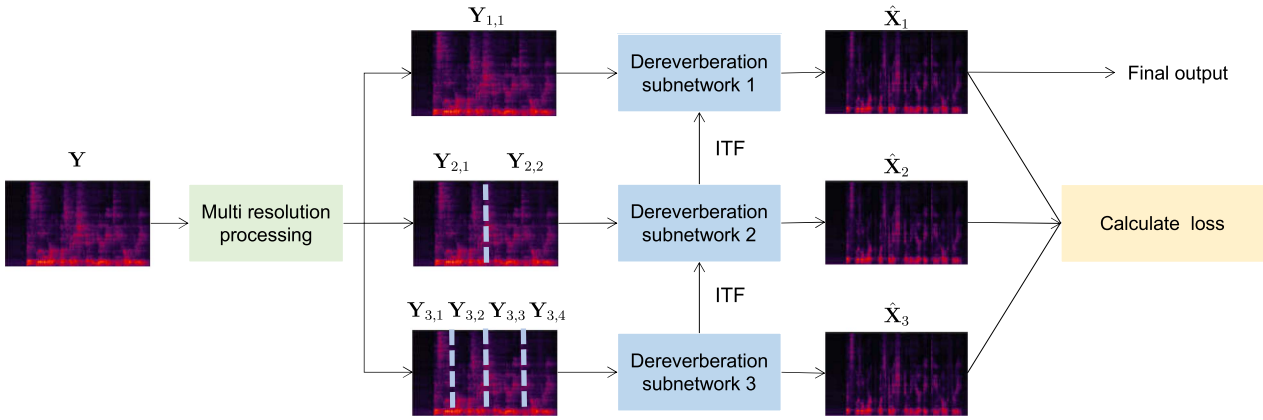


Fig. 1. Proposed multi-resolution framework.

the estimation of $\mathbf{x}$, denoted as $\hat{\mathbf{x}}$, we can get the estimated time-domain signal $\hat{x}[n]$ by inverse STFT using $|\hat{\mathbf{x}}|$ and $\angle\mathbf{y}$.

B. Multi-Resolution Dereverberation Framework

As shown in Fig. 1, the multi-resolution dereverberation framework contains M branches, each of which contains a dereverberation subnetwork. For the mth branch with $m = 1, \ldots, M$, the framework first partitions $\mathbf{Y} = [|\mathbf{y}_1|, \ldots, |\mathbf{y}_N|]$ into q^{m-1} equal-length non-overlapping segments, denoted as $\{\mathbf{Y}_{m,1}, \ldots, \mathbf{Y}_{m,i}, \ldots, \mathbf{Y}_{m,q^{m-1}}\}$, where $q \in \mathbb{N}$ is a resolution hyperparameter that is usually set to 2. Then, it takes each segment $\mathbf{Y}_{m,i}$ as an input to get an estimation of the magnitude spectrogram of the clean speech, denoted as $\hat{\mathbf{X}}_{m,i}$. The estimations of all segments are concatenated as $\hat{\mathbf{X}}_m = [\hat{\mathbf{X}}_{m,1}, \ldots, \hat{\mathbf{X}}_{m,i}, \ldots, \hat{\mathbf{X}}_{m,q^{m-1}}]$.

We propose to integrate all branches as a whole by transferring dereverberation information from high-resolution branches to low-resolution branches in sequence. In other words, the dereverberation information of the mth branch is sent to the $(m-1)$th branch so as to steadily increase the dereverberation capability of the latter:

$$\begin{aligned}
(\hat{\mathbf{X}}_M, \mathbf{I}_M) &= f_M(\mathbf{Y}_M) \\
(\hat{\mathbf{X}}_{M-1}, \mathbf{I}_{M-1}) &= f_{M-1}(\mathbf{Y}_{M-1}, \mathbf{I}_M) \\
&\vdots \\
(\hat{\mathbf{X}}_m, \mathbf{I}_m) &= f_m(\mathbf{Y}_m, \mathbf{I}_{m+1}) \\
&\vdots \\
(\hat{\mathbf{X}}_2, \mathbf{I}_2) &= f_2(\mathbf{Y}_2, \mathbf{I}_3) \\
\hat{\mathbf{X}}_1 &= f_1(\mathbf{Y}_1, \mathbf{I}_2)
\end{aligned} \tag{4}$$

where $f_m(\cdot)$ denotes the mth dereverberation branch, and $\mathbf{I}_{m+1}$ denotes the dereverberation information from the $(m+1)$th branch. After steadily increasing the dereverberation capability, the dereverberation output of the first branch, i.e. $\hat{\mathbf{X}}_1$, is used as the final output of the framework. Note that, $\mathbf{I}$ provides supplement local information for the low-resolution branches to improve their dereverberation performance. The multi-stage nature of the proposed model breaks down the challenging dereverberation task into sub-tasks, for progressively enhancing a distorted spectrogram. At the early stage, multiple non-overlapping segments provide multi-scale contextualized features for the dereverberation subnetwork; and the intermediate dereverberation output of each subnetwork plays like a reference for the next stage. At the final stage, the dereverberation subnetwork has abundant reference information provided from early stages, which helps the final stage delivers stronger output in any range of reverberation time than that of a single network.

The loss function of the proposed framework ℓ_{all} is defined as:

$$\ell_{\text{all}} = \sum_{m=1}^{M} w_m \ell_m(\hat{\mathbf{X}}_m, \mathbf{X}_m) \tag{5}$$

where ℓ_m is the loss function of the mth dereverberation network branch, $\mathbf{X}_m = [|\mathbf{x}|_1, \ldots, |\mathbf{x}|_N]$ is the magnitude spectrogram of the direct speech, and $w_m \in [0, 1]$ is the weight of the mth branch with a constraint $\sum_{m=1}^{M} w_m = 1$. The weights can be set manually or learned automatically. For simplicity, we set $w_m = 1/M, \forall m = 1, \ldots, M$.

From the above framework, we can see that the multi-resolution stacking in [32] is a special case of the proposed framework with $\mathbf{I}_m = \hat{\mathbf{X}}_m$. However, this early work does not jointly train the branches. In the following subsections, we aim to develop new implementations of the framework with recent advanced CRNNs, and further jointly train the branches with newly designed ITFs.

C. Implementation of the Framework Based on Convolutional Residual Neural Networks

As shown in Fig. 2, we focus on describing the mth branch with $1 < m \leq M$, which consists of a CB, a dereverberation subnetwork, an ITF, and a MB that are connected in sequence. All components are based on CRNNs. The only difference between the two implementations is that they use different dereverberation subnetworks. A general description of the implementations are as follows:

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.

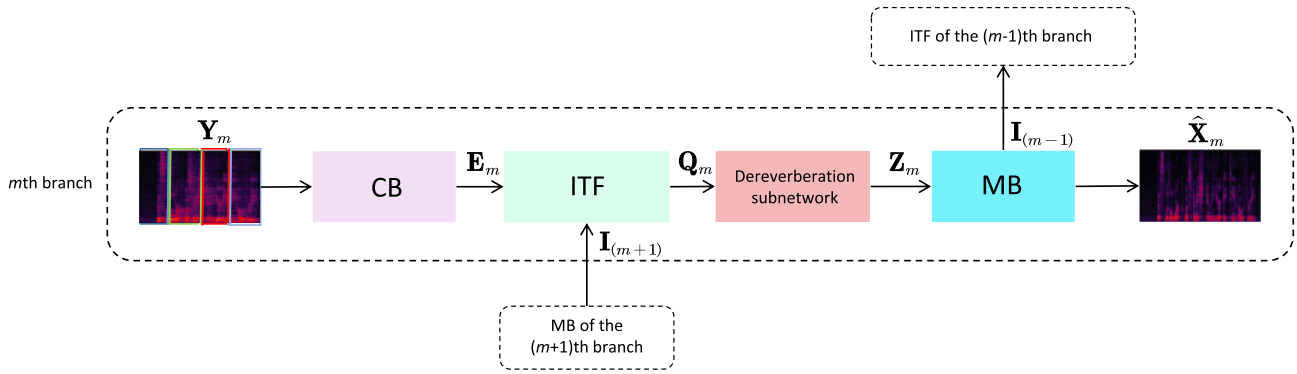


Fig. 2. Architecture of the proposed multi-resolution UNet.

The core contribution is that we integrate them properly into an effective system for our special speech dereverberation task. Specifically, CB, which was originally proposed as a building block for image super-resolution [39], transforms the original magnitude spectrogram of each input segment into C features by C convolutional channels. ITF, which was proposed originally for object detection [40], [41], fuses each of the feature with its corresponding intermediate dereverberation feature from the $(m+1)$th branch. Different from [40], [41], we just use 1×1 CBs to refine the intermediate dereverberation feature, and propagate them to the next stage for aggregation. The dereverberation subnetwork conducts dereverberation on each of the convolutional features. MB, which was proposed originally for image deblurring [42], fuses all C convolutional channels into a single output, and further produces C intermediate dereverberation features as part of the input of ITF in the $(m-1)$th branch. When $m=1$, because we do not need to generate intermediate dereverberation features by MB, we simply fuse all C channels by a convolution operator without resorting to a MB.

We use the spectral mapping [7], i.e. $\ell_m = \|\hat{\mathbf{X}}_m - \mathbf{X}_m\|_2$, as the training objective, though other advanced training objectives could be employed as well, such as complex ratio masking [38]. We present the components in detail as follows.

D. Convolutional Block

As shown in Fig. 3, CB first transforms each input segment $\mathbf{Y}_{m,i}$ into C features, denoted as $\{\mathbf{P}_{m,i,1},\ldots,\mathbf{P}_{m,i,c},\ldots\mathbf{P}_{m,i,C}\}$, by C convolutional channels, $\forall i=1,\ldots,q^{m-1}$.

Then, for each channel, CB transforms $\mathbf{P}_{m,i,c}$ through two convolutional layers and an activation function of the parametric rectified linear unit (PReLU) [43]:

$$\bar{\mathbf{P}}_{m,i,c} = \text{conv}_{3\times3}\left(\text{PReLU}\left(\text{conv}_{3\times3}\left(\mathbf{P}_{m,i,c}\right)\right)\right) \tag{6}$$

where $\text{conv}_{3\times3}(\cdot)$ is a 3×3 convolution operator. Next, a global pooling layer transforms $\bar{\mathbf{P}}_{m,i,c}$ to $p_{m,i,c}$ by:

$$p_{m,i,c} = \frac{1}{HW}\sum_{t=1}^{H}\sum_{f=1}^{W}\bar{P}_{m,i,c}(t,f) \tag{7}$$

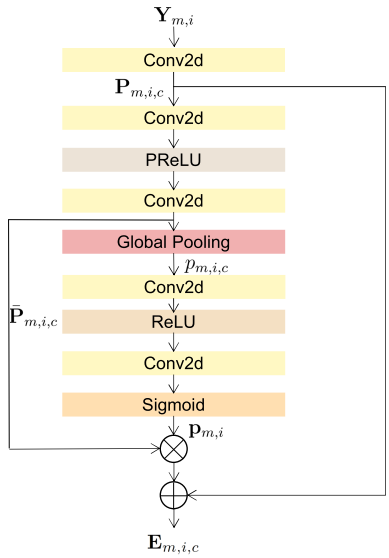


Fig. 3. Architecture of the CB.

where H and W represent the length and width of $\bar{\mathbf{P}}_{m,i,c}$ respectively, and $\bar{P}_{m,i,c}(t,f)$ is an element of $\bar{\mathbf{P}}_{m,i,c}$ at the tth column and fth row. Note that, the global pooling function is added to describe the global information of the spectrogram [44].

To exchange information across the convolutional channels, a simple gating mechanism with a sigmoid function is further added behind the global pooling function:

$$\bar{\mathbf{p}}_{m,i} = \text{sigmoid}((\mathbf{W}_2\text{ReLU}\left(\mathbf{W}_1\mathbf{p}_{m,i}\right))) \tag{8}$$

where

$$\mathbf{p}_{m,i} = [p_{m,i,1},\ldots,p_{m,i,C}]^T \tag{9}$$

and $\text{ReLU}(\cdot)$ denotes the rectified linear unit activation function, and $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weight matrices of the two convolutional layers whose sizes are $r\times C$ and $C\times r$ respectively, where r is a hyperparameter.

Finally, the output of CB is obtained by:

$$\mathbf{E}_{m,i,c} = \mathbf{P}_{m,i,c} + \bar{p}_{m,i,c}\bar{\mathbf{P}}_{m,i,c} \tag{10}$$

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.

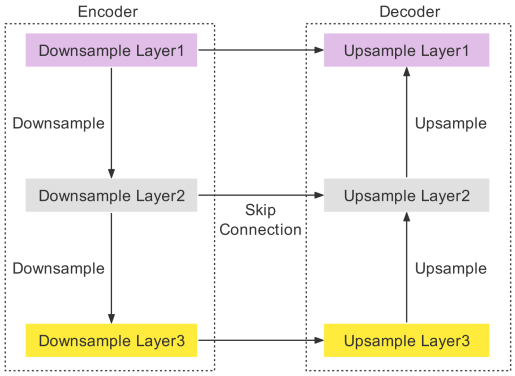


Fig. 4. Diagram of UNet.

where $\bar{p}_{m,i,c}$ is the cth element of $\bar{\mathbf{p}}_{m,i}$, and the summation operator between $\mathbf{P}_{m,i,c}$ and $\bar{p}_{m,i,c}\bar{\mathbf{P}}_{m,i,c}$ is a skip connection of CRNN for improving the robustness of the network training.

E. Information Transfer Between Dereverberation Branches

For each convolutional channel, ITF concatenates $\mathbf{E}_{m,i,c}$ with an intermediate dereverberation feature $[\mathbf{I}_{m+1,q(i-1)+1,c}, \ldots, \mathbf{I}_{m+1,qi,c}]$ produced by the MB of the $(m+1)$th branch:

$$\mathbf{Q}_{m,i,c} = \begin{bmatrix} \mathbf{E}_{m,i,c} \\ \text{conv}_{3\times 3}\left([\mathbf{I}_{m+1,q(i-1)+1,c}, \ldots, \mathbf{I}_{m+1,qi,c}]\right) \end{bmatrix} \tag{11}$$

where the intermediate dereverberation feature has the same time duration as $\mathbf{E}_{m,i,c}$.

F. Dereverberation Subnetworks

For each convolutional channel, a dereverberation subnetwork based on CRNN obtains a dereverberant feature $\mathbf{Z}_{m,i,c}$ from $\mathbf{Q}_{m,i,c}$. Here we first describe a UNet-based dereverberation subnetwork which has demonstrated its effectiveness on the dereverberation task, and then propose a CB-based dereverberation subnetwork.

1) Unet: As shown in Fig. 4, the UNet in a branch contains an encoder and a decoder. The encoder consists of three downsampling layers. The decoder consists of three upsampling layers.[1] There will be a downsampling operation to halve the feature dimension between the two downsampling layers and an upsampling operation to recover the feature dimension between the two upsampling layers. Each downsampling layer is connected to the corresponding upsampling layer by a skip connection. Particularly, each downsampling or upsampling layer is composed of 2 CB, which is to our knowledge a new implementation of UNet for speech dereverberation. Like [45], we use a bilinear upsampling followed by a convolutional layer as the upsampling layer, instead of using the transposed convolution. This modification reduces the checkerboard artifacts caused by the transposed convolution. The proposed implementation with UNet is denoted as MR-UNet.

[1]To reduce the large model size introduced by the increased number of branches, we only use three upsampling layers and three downsampling layers here.

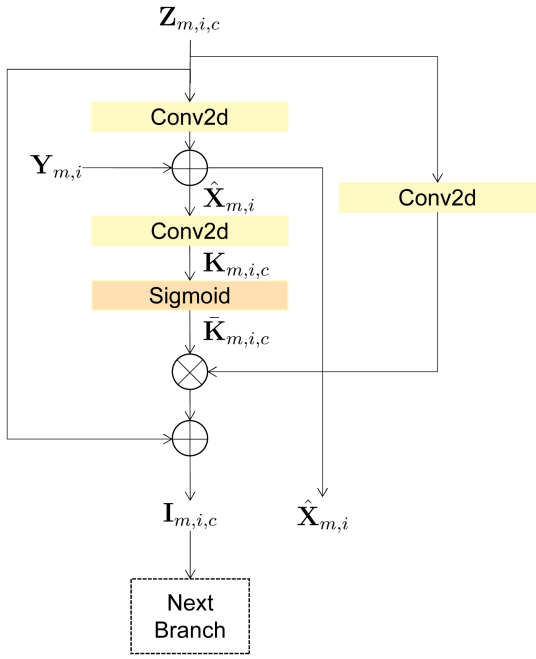


Fig. 5. Diagram of the MB.

2) Stacked Convolutional Blocks: To avoid possible information loss induced by repeating downsampling operations, we stack multiple CB into a new dereverberation network, denoted as stacked CB (SCB), followed by a convolutional layer. The input and output of SCB are further connected by a skip connection. The proposed implementation with SCB is denoted as MR-SCB.

G. Mask Block

As shown in Fig. 5, MB first fuses the intermediate dereverberation features of the dereverberation subnetworks, i.e. $\{\mathbf{Z}_{m,i,c}\}_{c=1}^{C}$, into a single feature by a 1×1 convolution operator, followed by a skip connection connected to the input magnitude spectrogram of the mth branch $\mathbf{Y}_{m,i}$:

$$\hat{\mathbf{X}}_{m,i} = \mathbf{Y}_{m,i} + \text{conv}_{1\times 1}(\mathbf{Z}_{m,i,1}, \ldots, \mathbf{Z}_{m,i,C}). \tag{12}$$

Then, MB produces $\{\mathbf{I}_{m,i,c}\}_{c=1}^{C}$ from $\hat{\mathbf{X}}_{m,i}$ for the $(m-1)$th branch. Specifically, it first transforms $\hat{\mathbf{X}}_{m,i}$ into C features by a 1×1 convolution operator, denoted as $\{\mathbf{K}_{m,i,1}, \ldots, \mathbf{K}_{m,i,c}, \ldots, \mathbf{K}_{m,i,C}\}$, and then gets a set of masks by a sigmoid activation function,

$$\bar{\mathbf{K}}_{m,i,c} = \text{sigmoid}(\mathbf{K}_{m,i,c}) \tag{13}$$

which are finally masked to the output of the dereverberation subnetwork:

$$\mathbf{I}_{m,i,c} = \text{conv}_{1\times 1}(\mathbf{Z}_{m,i,c})\bar{\mathbf{K}}_{m,i,c} + \mathbf{Z}_{m,i,c} \tag{14}$$

where the summation operator is a skip connection of CRNN.

IV. Experimental Setup

This section presents the experimental settings, including the datasets, comparison methods, and evaluation metrics.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.

A. Datasets

The evaluation was conducted on one simulated data and two real-world data. For the simulated data, we randomly selected 7000, 4000, and 7000 clean utterances from the Librispeech corpus [46] as the clean speech source of training, validation and test data, respectively. Specifically, the training and validation data are selected from the train-clean-100, train-clean-360 and train-other-500 of Librispeech, while the testing data are selected from the dev-clean, dev-other, test-clean and test-other of Librispeech. The room impulse response (RIR) function was generated by the image source model [47]. For each utterance, we generated a room with its length, width, and height randomly selected from [3,10], [3,8], and $[2.5, 6]$ meters respectively. The reverberation time T_{60} of the room was randomly generated from $[0.2, 1.2]$ seconds. The speech source and a microphone were placed randomly in the room and satisfy the following constraints: (i) the distance between them was controlled to be in [0.5,10] meters, and (ii) the distance from the speech source or microphone to the walls was controlled to be at least 0.3 meters.

The first real-world data was the real recorded reverberant utterances from the Libri-adhoc40 corpus [48]. It is a replayed version of the Librispeech corpus in a real office environment. The recording environment is an office room with a size of $9.8 \times 10.3 \times 4.2$ meters. The room is highly reverberant with T_{60} around 0.9 s and little additive noise. Each replayed utterance has 40 recordings recorded by 40 microphones. The locations of both the microphones and speakers of the training, evaluation, and test data are different. The distances between the speakers and the microphones were ranged from 0.8 m to 7.4 meters, which makes the dataset suitable for the study of far-field speech processing. For each replayed utterance, we randomly selected one reverberant recording from all 40 recordings. Finally, we had 28540 training utterances, 2621 testing utterances and 2704 validating utterances. Moreover, the clean utterances of Libri-adhoc40, which are replayed recordings of Librispeech in a full anechoic chamber, was used for model training and evaluation. They were recorded by the same microphones and speakers as those in the office room, so as to eliminate the effect of the equipments.

The second real-world data was the real recorded reverberant utterances from the VOICES corpus [49], which is a replayed version of the Librispeech corpus in acoustically challenging conditions. The recordings took place in four rooms of various sizes, each of which has its own background and reverberation profile. Four types of distractor noise were simultaneously played with clean speech. Here, we randomly selected 25560 training utterances and 6400 validation utterances from the train subset of the VOICES corpus, and 6400 testing utterances from its test subset.

B. Comparison Methods

Our network was trained for 100 epochs with batch size 4. We used Adam optimizer [50] with β_1 and β_2 set to 0.9 and 0.999 respectively. The initial learning rate is 2×10^{-4}, which is steadily decreased to 1×10^{-6} using the cosine annealing strategy [51] and a warmup step of 5250. Our MR-UNet contains 3 branches. The CB of each branch transforms the input acoustic feature into 96 channels. The dimension r of $\mathbf{W}_1$ and $\mathbf{W}_2$ in (8) was set to 6. Each branch of the UNet contains 3 downsampling layers and 3 upsampling layers, where each layer consists of 2 CB. The downsampling rate between two downsampling layers is 0.5. The upsampling rate between two upsampling layers is 2. The first downsampling layer has 96 channels. The second downsampling layer has 144 channels. The third downsampling layer has 192 channels. The number of channels of each upsampling layer is the same as the number of the channels of its corresponding downsampling layer. Our proposed MR-UNet contains around 25.78 M parameters. For the proposed MR-SCB, each SCB dereverberation network consists of a stack of 8 CB.

We compared the proposed method with the following representative dereverberation algorithms [2]:

- *Weighted prediction error (WPE):* It is a well-known conventional dereverberation algorithm. Here we use the open source NARA-WPE algorithm [28] as an implementation.
- *Long short-term memory (LSTM) [10]:* It concatenates the magnitude spectrograms of neighboring frames in a context window as its input, and predicts the central frame of the window. The window size is 11. The LSTM model consists of two LSTM layers with 400 hidden units per layer. The number of parameters of the model is around 6.13 M.
- *Late reverberation suppression LSTM (Late-LSTM) [11]:* It first estimates late reverberation by an LSTM model, and then subtracts the estimated late reverberation from the magnitude spectrogram of the reverberant speech. It consists of two LSTM hidden layers with 512 hidden units per layer. The dropout rate of the LSTM layers was set to 0.3. The number of parameters of the model is around 3.48 M.
- *UNet [13]:* It uses an encoder-decoder network with skip connections. The encoder contains four downsampling layers, each of which downsamples its input with a stride of 2 to the next layer until there is a bottleneck. The decoder conducts a reverse process by upsampling its input until the output reaches to the original size of the input of the network. The number of the parameters of the model is around 31.04 M.
- *Skip convolutional neural network (SkipConvNet) [15]:* It has a similar structure with UNet, which consists of an encoder of 8 downsamling layers and a decoder of 8 upsampling layers. The skip connection between the encoder layer and its corresponding decoder layer is replaced by multiple convolutional modules, which improve the learning capacity of the network by providing the decoder with intuitive feature maps rather than the output of the encoder. The number of parameters of the model is around 64.33 M.

[2]The source codes of all baselines are available on at: https://github.com/fgnt/nara_wpe (WPE), https://github.com/DiegoLeon96/Neural-Speech-Dereverberation(LSTM, Late LSTM and UNet), https://github.com/zehuachenImperial/SkipConvNet (SkipConvNet), https://github.com/key2miao/CAUNet (CAUnet), https://github.com/Andong-Li-speech/DARCN (DARCN), https://github.com/sp-uhh/sgmse (SGMSE and SGMSE+).

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.

- *CAUnet [23]:* It is also a UNet-based network, which uses the dilated-dense block in each layer of the encoder and decoder. It adopts stacked two-stage transformer blocks between the encoder and decoder to extract contextual information from the output of the encoder. The number of parameters of the model is around 1.04 M.
- *DARCN [29]:* It is a separated sub-network, which adaptively generates an attention distribution to control the information flow throughout the major network. By introducing recursive learning, the number of its trainable parameters is reduced dynamically, because of reusing a network for multiple stages. The number of parameters of the model is around 1.23 M.
- *SGMSE [30]:* It is a score-based generative models, which adopts DCUNet [17] as the backbone network. So called time-embedding layers, which aim to provide the information of the time steps into the network, are added into the encoder and decoder blocks. The network has two input channels for noisy and clean spectrograms respectively, and one output channel for the prediction. The number of parameters of the model is around 3.5 M.
- *SGMSE+ [31]:* It is an improved version of SGMSE, where its DCUNet is replaced by the Noise Conditional Score Network, a more sophisticated than the former. The number of parameters of the model is around 65.6 M.

C. Evaluation Metrics

We used short-time objective intelligibility (STOI) [52], perceptual evaluation of speech quality (PESQ) [53], and frequency-weighted segmental signal-to-noise ratio (fwSegSNR) [54] to evaluate the dereverberation performance at the signal level. PESQ is a phase-aware metric for speech quality. STOI evaluates the objective intelligibility of a degraded speech signal by computing the correlation of the temporal envelopes of the degraded speech signal and its clean reference. FwSegSNR evaluates the signal discrepancy between dereverberant speech and clean speech in a reweighted frequency domain where the weights of the frequency bins are calculated based on the amplitude of the clean speech.

V. Experimental Results on Simulated Data

This section first reports the results of the proposed MR-UNet and 9 referenced methods on the simulated data in general, then evaluates the components of proposed method in detail. Because the state-of-the-art speech dereverberation models are based on UNet, we focus on presenting the result of the proposed MR-UNet in Sections V-A to V-B, so as to better demonstrate the advantage of the proposed method, leaving the result of the proposed MR-SCB in Section V-C as a supplemental discussion to MR-UNet.

A. Main Results

Table I lists the results of the comparison methods on the simulated data. From the table, we observe that, compared to the reverberant speech, most dereverberation algorithms except the LSTM dereverberation algorithm improves the speech quality significantly. For example, the proposed MR-UNet improves the speech quality over the unprocessed reverberant speech by 18.4% in STOI, 0.61 in PESQ, and 4.65 dB in fwSegSNR. For the comparison methods, the proposed MR-UNet outperforms the referenced methods in all metrics. For example, it achieves relative STOI improvement of 17.7% and 2.02 dB fwSegSNR improvement over the runner-up method UNet, and relative PESQ improvement of 8.7% over the runner-up method SkipConvNet. Note that, the overall performance of the UNet and its variants is better than that of the LSTM and its variants, while WPE does not work well in this highly reverberant scenario.

TABLE I
Results of the Comparison Methods on the Simulated Dataset

Test data	Method	Model size	STOI	PESQ	fwSegSNR
Simulated	Reverberant speech	-	0.693	2.006	6.576
	WPE [28]	-	0.708	2.038	6.634
	LSTM [10]	6.13 M	0.461	2.121	5.683
	Late-LSTM [11]	3.48 M	0.724	2.020	7.304
	UNet [14]	31.04 M	0.842	2.353	9.203
	SkipConvNet [15]	64.33 M	0.826	2.435	8.979
	CAUnet [23]	1.04 M	0.797	2.208	8.222
	DARCN [29]	1.23 M	0.814	2.280	8.420
	SGMSE [30]	3.50 M	0.720	1.938	6.752
	SGMSE+ [31]	65.60 M	0.819	2.414	8.905
	MR-UNet (proposed)	25.78 M	**0.877**	**2.616**	**11.226**

The number in bold indicates the best performance.

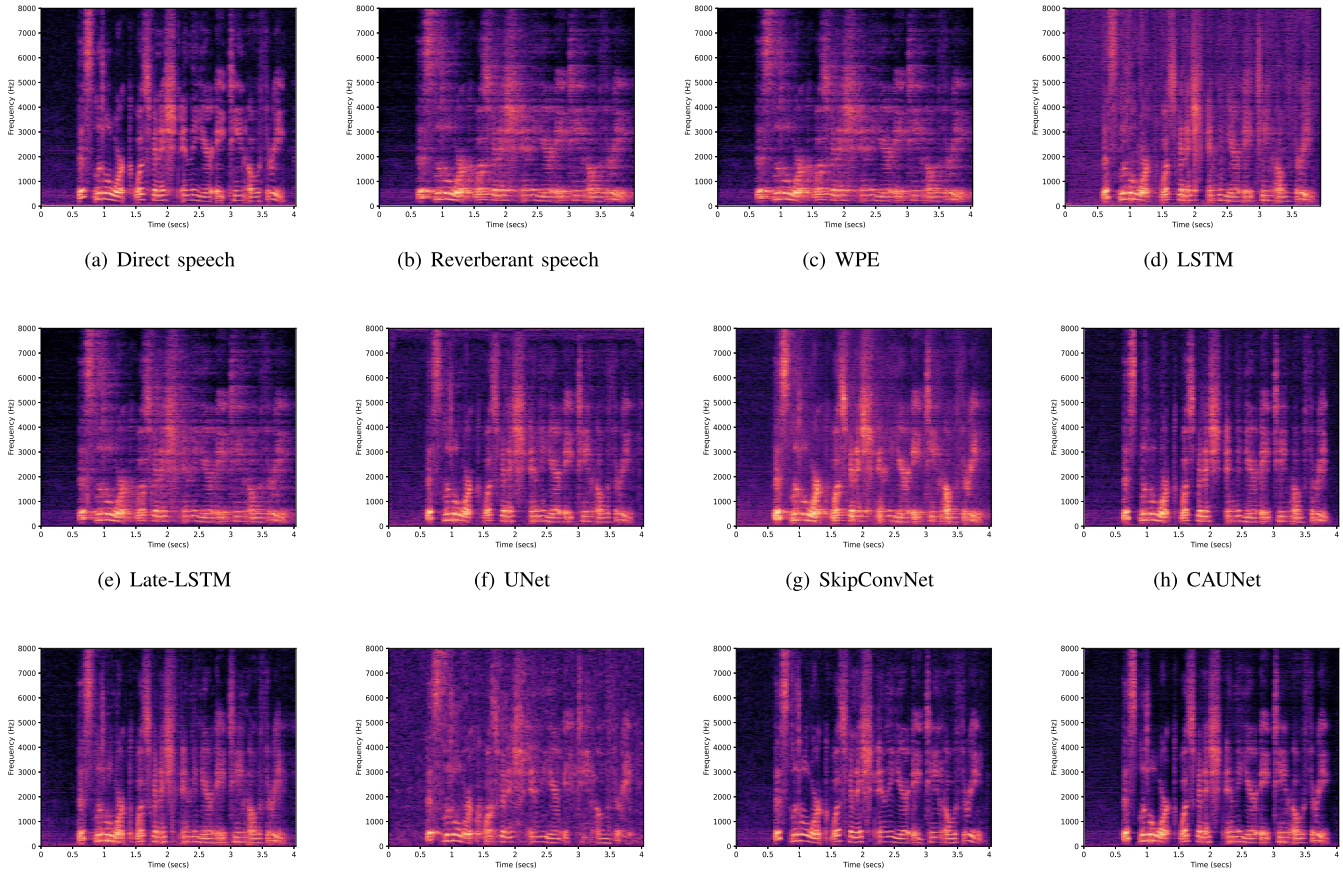
Fig. 6 visualizes the magnitude spectrograms of the dereverberant speech produced by the comparison methods on a random sample with $T_{60} = 756$ ms of the simulated dataset. From the figure, we see that the proposed MR-UNet performs well in suppressing the smearing effect caused by the reverberation. It behaves similarly with SkipConvNet and SGMSE+ in general, and seems better than the latter two methods in maintaining a clear local pattern of the speech. It achieves apparently better performance than the remaining comparison methods. Note that, although UNet achieves runner-up in STOI of Table I, it suffers from a spectrogram leakage problem. In addition, although LSTM maintains the spectrogram pattern clearly, it introduces unexpected artifacts.

B. Ablation Study

This subsection studies the effects of the number of branches and the ITF of MR-UNet on performance, as well as the effect of MR-UNet on different reverberation time.

1) Effects of the Number of Branches and the Information Transfer Function: In this subsection, we study the effect of the number of branches of MR-UNet, particularly with or without CB. Specifically, in Fig. 2, CBs are used in two positions of the proposed multi-resolution algorithm: (i) CB1 is located between the input image and ITF, and (ii) CB2s are the components of the dereverberation subnetwork. As mentioned in Section III-F, both subnetworks, i.e. UNet and SCB, are composed of CBs. Here, to address whether the performance improvement of the MR-UNet with respect to the number of branches is related to CB, we use plain convolutional layers instead of CBs, and study the performance of MR-UNet in the absence of CBs.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.



(a) Direct speech (b) Reverberant speech (c) WPE (d) LSTM

(e) Late-LSTM (f) UNet (g) SkipConvNet (h) CAUNet

(i) DARCN (j) SGMSE (k) SGMSE+ (l) MR-UNet

Fig. 6. Magnitude spectrograms of the unprocessed reverberant speech and the dereverberant speech produced by the comparison methods on a random simulated sample with $T_{60} = 756$ ms.

TABLE II
EFFECT OF THE NUMBER OF BRANCHES AND CBS OF MR-UNET ON PERFORMANCE, WHERE M DENOTES THE NUMBER OF BRANCHES OF MR-UNET

Test data	M	CB1	CB2	STOI	PESQ	fwSegSNR
Simulated	1	✗	✗	0.761	2.054	7.372
	2	✗	✗	0.794	2.172	7.898
	3	✗	✗	0.817	2.305	8.669
	4	✗	✗	0.825	2.332	8.970
	4	✓	✗	0.845	2.438	9.591
	4	✓	✓	**0.882**	**2.629**	**11.220**

TABLE III
EFFECT OF THE ITFS OF MR-UNET WITH $M = 3$ ON PERFORMANCE. THE TERM "ITF3-2" INDICATES THE ITF FROM BRANCH 3 TO BRANCH 2, WHILE "ITF2-1" INDICATES THE ITF FROM BRANCH 2 TO BRANCH 1

Test data	Dereverberation subnetworks	ITF3-2	ITF2-1	STOI	PESQ	fwSegSNR
Simulated	UNet in branch 3	✗	✗	0.863	2.550	10.533
	UNet in branch 2	✗	✗	0.866	2.554	10.545
		✓	✗	**0.874**	**2.606**	**11.079**
	UNet in branch 1	✗	✗	0.863	2.557	10.566
		✓	✗	0.858	2.561	10.541
		✗	✓	0.875	2.599	11.083
		✓	✓	**0.877**	**2.616**	**11.226**

The symbol "✗" means that the corresponding ITF was disabled by setting its output to zero.

Table II investigated the effect of the number of branches and CBs of MR-UNet, where we increased M from 1 to 4. From the table, generally we observe that the performance of MR-UNet without CB is increasing with respect to the number of branches. Specifically, the MR-UNet with four branches outperforms that with a single branch by relatively 26.7% in STOI, relatively 11.4% in PESQ, and 1.59 dB in fwSegSNR. We also can see that the MR-UNet with CB1 and CB2 further outperforms that without CB by relatively 32.5% in STOI, relatively 13.7% in PESQ, and 2.25 dB in fwSegSNR, which verifies the effectiveness of CB.

To further investigate the influence of the information transfer functions on performance, we list the results of all three dereverberation subnetworks of MR-UNet in Table III. We observe that, without ITF3-2, the performance of the UNet subnetwork in branch 2 degrades to the same result as the UNet subnetwork in branch 3. For the UNet in branch 1, the same result appears when we turn off both ITF3-2 and ITF2-1. The UNet in branch 1 achieves the best performance only when both ITFs are activated.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
IMPACT OF THE INFORMATION FLOW DIRECTION ON PERFORMANCE. FORWARD: THE PROPOSED MR-UNET

Test data	Model	STOI	PESQ	fwSegSNR
Simulated	Forward	0.877	2.616	11.226
	Reverse	0.880	2.605	11.134

Reverse: information flows in the reverse direction of the forward model, i.e. from branch number 1 to 3.

TABLE V
COMPARISON BETWEEN MR-UNET THAT HAS MULTIPLE BRANCHES AND A MODIFIED MR-UNET WITH ONLY A SINGLE BRANCH AND ENLARGED MODEL SIZE

Test data	Model name	Model size	STOI	PESQ	fwSegSNR
Simulated	MR-UNet, $M = 1$	17.91 M	0.869	2.568	10.751
	Modified MR-UNet, $M = 1$	29.80 M	0.873	2.587	10.781
	MR-UNet, $M = 3$	25.78 M	**0.877**	**2.616**	**11.226**

All the above results verified the effectiveness of the ITFs in the proposed method.

To study whether the information flow direction will affect the effectiveness of the proposed method, we take the standard MR-UNet whose information flow direction is from branch 3 to branch 1 as the Forward model, and constructed another MR-UNet whose information flows from branch 1 to branch 3 as the Reverse model. Table IV lists the performance comparison. From the table, it can be seen that the Reverse model yields similar results to the Forward model, indicating that information can be transmitted in different directions. This also meets our expectations. Because each branch is only responsible for learning spectral information at different scales, the order of the branches is not important.

To study whether the improvement is caused by simply increasing the number of parameters, we constructed a MR-UNet with a single branch, and increased its parameters to 29 M by adding two dowmsample layers and two upsample layers. Eventually, the number of its parameters is comparable to the MR-UNet with three branches. We compared two MR-UNets ($M = 1$ and $M = 3$) with the modified one-branch MR-UNet (Modified MR-UNet, $M = 1$) in Table V. From the table we observe that increasing the model size improves the performance of the Modified MR-UNet over "MR-UNet ($M = 1$)", but there is sill a gap between Modified MR-UNet and "MR-UNet ($M = 3$)", which also proves the effectiveness of the MR-UNet and particularly the information transfer function.

2) Effect on Different Reverberation Time: To study how the robustness of the proposed algorithm has improved against different reverberation time when the number of branches is increased, we divided the test data into 5 groups according to the reverberation time, and evaluated the performance of the algorithm on each group. Here we first define two evaluation metrics ΔSTOI and ΔPESQ. ΔSTOI indicates the absolute value of the difference of the STOI scores between the proposed MR-UNet and the original reverberant speech. A similar definition applies to ΔPESQ too.

Table VI shows the statistical results. From the table, we observe that the values of ΔSTOI and ΔPESQ tends to stabilize when the number of branches increases. That is to say, with the increased number of branches, the difference of the scores (ΔSTOI or ΔPESQ) between each pair in a group will be decreasing, which indicates that our proposed multi-resolution method can improve the robustness against different reverberation times.

TABLE VI
IMPROVEMENT OF THE PROPOSED MR-UNET OVER THE ORIGINAL REVERBERANT SPEECH IN DIFFERENT GROUPS OF TEST DATA ACCORDING TO THE REVERBERATION TIME

Test data	M	T_{60}	ΔSTOI	ΔPESQ
Simulated	1	0.2s-0.4s	0.145	0.579
		0.4s-0.6s	0.171	0.604
		0.6s-0.8s	0.180	0.589
		0.8s-1.0s	0.180	0.557
		1.0s-1.2s	0.176	0.518
	2	0.2s-0.4s	0.159	0.584
		0.4s-0.6s	0.182	0.628
		0.6s-0.8s	0.191	0.616
		0.8s-1.0s	0.191	0.591
		1.0s-1.2s	0.188	0.562
	3	0.2s-0.4s	0.160	0.653
		0.4s-0.6s	0.183	0.676
		0.6s-0.8s	0.192	0.655
		0.8s-1.0s	0.192	0.626
		1.0s-1.2s	0.189	0.593
	4	0.2s-0.4s	0.162	0.633
		0.4s-0.6s	0.184	0.658
		0.6s-0.8s	0.193	0.643
		0.8s-1.0s	0.193	0.618
		1.0s-1.2s	0.189	0.586

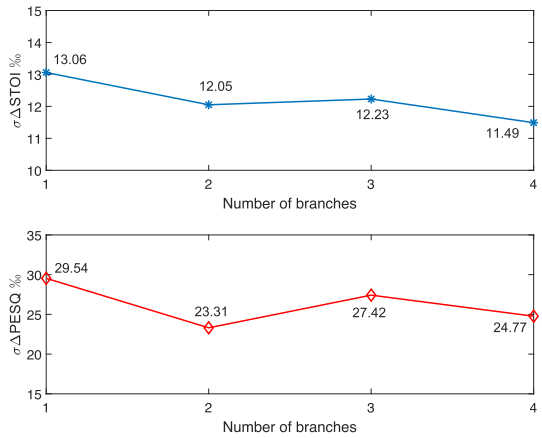


Fig. 7. Curves of $\sigma\Delta$STOI and $\sigma\Delta$PESQ with respect to the number of branches of MR-UNet.

To illustrate Table VI more clearly, we further define two evaluation metrics $\sigma\Delta$STOI and $\sigma\Delta$PESQ: $\sigma\Delta$STOI represents the standard deviation of the ΔSTOI scores in a group of test data, and $\sigma\Delta$PESQ is defined similarly.

Fig. 7 plots the $\sigma\Delta$STOI and $\sigma\Delta$PESQ scores for different number of branches. From Fig. 7, we observe a decrease trend in terms of both $\sigma\Delta$STOI and $\sigma\Delta$PESQ as the number of branches increases. Specifically, $\sigma\Delta\text{STOI} = 13.06$ and $\sigma\Delta\text{PESQ} = 29.54$ when branch $= 1$, and they drop to 11.40 and 24.77 respectively when branch $= 4$. The results support our conclusion that the multi-resolution approach is effective in improving the robustness of the system against different reverberation time when the number of branches increases.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
COMPARISON BETWEEN MR-UNET AND MR-SCB

Test data	M	Model Name	Model size	STOI	PESQ	fwSegSNR
Simulated	1	MR-SCB	4.68 M	**0.871**	2.565	10.571
		MR-UNet	17.91 M	0.869	**2.568**	**10.751**
	2	MR-SCB	9.53 M	0.874	**2.602**	11.052
		MR-UNet	20.06 M	**0.875**	2.599	**11.060**
	3	MR-SCB	14.38 M	**0.882**	**2.640**	11.156
		MR-UNet	25.78 M	0.877	2.616	**11.226**
	4	MR-SCB	19.24 M	0.879	2.616	11.167
		MR-UNet	31.84 M	**0.882**	**2.629**	**11.221**

TABLE VIII
RESULTS OF THE COMPARISON METHODS ON THE REAL-WORLD LIBRI-ADHOC40 DATASET

Test data	Methods	STOI	PESQ	fwSegSNR
Libri-adhoc40	Reverberant speech	0.539	1.645	4.218
	WPE [28]	0.548	1.651	4.249
	LSTM [10]	0.252	1.301	2.349
	Late-LSTM [11]	0.573	1.641	5.320
	UNet [14]	0.702	1.817	6.553
	SkipConvNet [15]	0.692	**2.226**	7.578
	CAUNet [23]	0.634	1.771	6.446
	DARCN [29]	0.651	1.630	4.313
	SGMSE [30]	0.319	1.128	3.492
	SGMSE+ [31]	0.699	2.168	7.373
	MR-SCB	0.728	2.201	8.211
	MR-UNet	**0.752**	2.223	**8.768**

C. Result of Multi-Resolution Stacked Convolutional Blocks

To study the effectiveness of the proposed SCB, we compare MR-UNet with MR-SCB in the setting of four branches. In Table VII, we observe that the performance of both models is improved with respect to the number of branches, at the expense of increased complexity. This trend is consistent with the result in Table II. In addition, comparing Tables I and VII, we observe that the MR-UNet with a single branch, which degrades to our novel UNet implementation based on CB, outperforms the baseline UNet. This phenomenon further demonstrates the advantage of our UNet implementation. We also see that MR-SCB and MR-UNet have similar performance in both STOI and PESQ, and the latter has higher fwSegSNR scores in all scenarios. Moreover, although they yield similar performance, the proposed MR-SCB can use less parameters than MR-UNet. The most effective improvement for both models happens when the number of branches M from 1 to 2. When M is increased from 3 to 4, the performance of both models is increased slightly while the number of parameters is increased by 23% for MR-UNet and 33.8% for MR-SCB. Eventually, we set $M = 3$ for both models in the following experiments so as to balance the model size and performance.

VI. EXPERIMENTAL RESULTS ON REAL-WORLD DATA

This section studies the dereverberation performance of the comparison methods on the two real-world data.

Table VIII lists the results of the comparison methods on the real-world Libri-adhoc40 dataset. From the table, we see that the proposed MR-UNet and MR-SCB achieve the best performance among the comparison methods in the strongly-reverberant real-world scenario. For example, they achieve 22% and 16.6% relative improvement respectively over the best referenced method UNet in STOI, 1.88 dB and 1.47 dB improvement respectively over the best referenced method SkipConvNet in fwSegSNR. The proposed MR-UNet behaves similarly to SkipConvNet in PESQ, while MR-SCB is slightly worse than SkipConvNet in PESQ. Note that, the UNet-based methods perform better than the LSTM-based methods, which is consistent with our observation in the simulation scenario.

TABLE IX
RESULTS OF THE COMPARISON METHODS ON THE REAL-WORLD VOICES DATASET

Test data	Methods	STOI	PESQ	fwSegSNR
VOICES	Reverberant speech	0.485	1.892	3.325
	WPE [28]	0.498	1.947	3.370
	LSTM [10]	0.462	0.962	2.443
	Late-LSTM [11]	0.517	1.692	3.901
	UNet [14]	0.746	1.925	6.977
	SkipConvNet [15]	0.739	2.042	6.707
	CAUNet [23]	0.743	2.071	6.513
	DARCN [29]	0.791	1.794	4.141
	SGMSE [30]	0.501	1.279	4.082
	SGMSE+ [31]	0.636	2.065	6.296
	MR-SCB	0.788	2.073	8.173
	MR-UNet	**0.802**	**2.094**	**8.594**

Table IX lists the results of the comparison methods on the real-world VOICES dataset. From the table, we observe that the proposed MR-UNet outperforms the referenced methods in all metrics. The proposed MR-SCB achieves the runner-up performance in PESQ and fwSegSNR, and is only slightly worse than MR-UNet and DARCN in STOI.

VII. APPLICATION TO FAR-FIELD SPEECH RECOGNITION

It is known that some deep learning based speech denoising and dereverberation methods may introduce strong distortion to the enhanced speech, which hinders their applications to speech recognition. For this problem, here we study the comparison methods on two speech recognition systems.

The acoustic model of the first speech recognition system for the evaluation is a conformer [55] trained with 960 hours of annotated speech from the Librispeech data. The language model is a transformer [56] trained using the transcripts of the 960 hours of annotated speech of Librispeech added with a text-only corpus of additional 800 M word tokens. The decoding algorithm is the joint CTC-attention decoding [57].

The architecture of the second speech recognition system is the same as the first one. Its acoustic model was trained with 4000 hours reverberant data and 100 hours clean data from Libri-adhoc40. Its language model was trained using the transcripts of the above 4100 hours of annotated speech added with the 800 M word token corpus. We used word error rate (WER) as the evaluation metric and took the dereverberant speech produced by the comparison methods on the real-world Libri-adhoc40 dataset as the input of the two ASR systems.

Table X lists the WERs of the comparison methods on the first ASR system. For the integrity of the table, we also reported the ASR system on the clean speech as a reference. From the table, we observe that the proposed MR-UNet and MR-SCB achieve the lowest WERs, which are 37.2% and 34.9% relatively lower than that of the best referenced method—SkipConvNet. Comparing Tables VIII and X, we see a positive correlation

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.

TABLE X
WERS OF THE COMPARISON DEREVERBERATION ALGORITHMS ON THE FIRST ASR SYSTEM WHICH WAS TRAINED WITH CLEAN SPEECH

Training set of ASR	Methods	WER (%)
Librispeech	Reverberant speech	60.96
	WPE [28]	57.25
	LSTM [10]	97.33
	Late-LSTM [11]	47.45
	UNet [14]	29.84
	SkipConvNet [15]	17.35
	CAUNet [23]	49.61
	DARCN [29]	43.51
	SGMSE [30]	95.72
	SGMSE+ [31]	25.43
	MR-SCB	11.29
	MR-UNet	**10.89**
	Original speech	*2.25*

TABLE XI
WERS OF THE COMPARISON DEREVERBERATION ALGORITHMS ON THE SECOND ASR SYSTEM WHICH WAS TRAINED WITH BOTH CLEAN SPEECH AND REVERBERANT SPEECH

Training set of ASR	Methods	WER (%)
Libri-adhoc40	Reverberant speech	**12.49**
	WPE [28]	17.55
	LSTM [10]	90.62
	Late-LSTM [11]	38.16
	UNet [14]	34.39
	SkipConvNet [15]	20.88
	CAUNet [23]	46.17
	DARCN [29]	41.73
	SGMSE [30]	87.33
	SGMSE+ [31]	22.68
	MR-SCB	16.43
	MR-UNet	15.86
	Original speech	*7.0*

between the dereverberation performance and WER for both conventional WPE and the deep learning based methods.

Table XI lists the WERs of the comparison methods on the second ASR system. For each test utterance, we not only give the ASR performance on the clean speech, but we also provide the reverberant speech from a random channel of all 40 channels. From the table, we see that, similar to the results in Table X, the proposed MR-SCB and MR-UNet also have apparent superiority over the other dereverberant methods. However, they are less effective than simply applying reverberant speech to the ASR. The phenomena was caused by that the ASR system was trained to fit the reverberant environment, therefore, the dereverberant speech produced from the comparison methods mismatches with the reverberant training data.

However, if we take Tables X and XI together, we find that the proposed method achieves the lowest WER. Moreover, because training the first ASR system with the clean speech is easier than training the second system with the noisy data that is over 40 times larger than the clean speech, the advantage of the proposed method was further demonstrated.

VIII. CONCLUSION

In this paper, we have proposed a multi-resolution framework to address speech dereverberation, and further proposed an implementation of the framework based on UNet. Specifically, the framework consists of multiple dereverberation branches, each of which partitions the input into non-overlapping segments under a specific resolution. The dereverberation branches are connected via information transfer functions, which help the branches jointly trained. We also proposed two implementations based on CRNN. The first one, named MR-UNet, is based on our new implementation of UNet based CB. The second one, named MR-SCB, is based on our newly proposed SCB dereverberation network. We have conducted a systematic comparison with 9 representative dereverberation methods in both a simulated environment and two real-world scenarios. Experimental results demonstrate that the proposed method outperforms the referenced methods in both of the environments. The experimental conclusion is consistent not only in the dereverberant effect but also in the application to far-field speech recognition. We also find that the performance of the proposed SCB in the multi-resolution framework is close to that of UNet in all experimental results.

Some weaknesses of the proposed method, such as the inefficiency for the on-line processing and for the processing of very long mixtures at the inference time, need further investigation in the future. Some improvement can also be made in the future, e.g. designing new training objectives in the time domain or complex domain.

REFERENCES

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.
[2] K. Lebart, J.-M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica United Acustica*, vol. 87, no. 3, pp. 359–366, 2001.
[3] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 774–784, May 2006.
[4] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.
[5] B. Schwartz, S. Gannot, and E. A. P. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 2, pp. 394–406, Feb. 2015.
[6] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 4628–4632.
[7] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.
[8] Y. Zhao, Z.-Q. Wang, and D. Wang, "A two-stage algorithm for noisy and reverberant speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5580–5584.
[9] J. F. Santos and T. H. Falk, "Speech dereverberation with context-aware recurrent neural networks," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 26, no. 7, pp. 1236–1246, Jul. 2018.
[10] M. Mimura, S. Sakai, and T. Kawahara, "Speech dereverberation using long short-term memory," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2435–2439.
[11] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Late reverberation suppression using recurrent neural networks with long short-term memory," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5434–5438.
[12] X. Tang, J. Du, L. Chai, Y. Wang, Q. Wang, and C.-H. Lee, "A LSTM-based joint progressive learning framework for simultaneous speech dereverberation and denoising," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2019, pp. 274–278.
[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.

[14] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech dereverberation using fully convolutional networks," in *Proc. IEEE 26th Eur. Signal Process. Conf.*, 2018, pp. 390–394.
[15] V. Kothapally, W. Xia, S. Ghorbani, J. H. Hansen, W. Xue, and J. Huang, "SkipConvNet: Skip convolutional neural network for speech dereverberation using optimally smoothed spectral mapping," in *Proc. INTERSPEECH*, 2020, pp. 3935–3939.
[16] N. Zheng and X.-L. Zhang, "Phase-aware speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 27, no. 1, pp. 63–76, Jan. 2019.
[17] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 7, pp. 1492–1501, Jul. 2017.
[18] J. Zhang, M. D. Plumbley, and W. Wang, "Weighted magnitude-phase loss for speech dereverberation," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 5794–5798.
[19] 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," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–20.
[20] Y. Hu et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.
[21] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation," in *Proc. 19th Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 334–340.
[22] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
[23] K. Wang, B. He, and W. Zhu, "CAUNet: Context-aware U-net for speech enhancement in time domain," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2021, pp. 1–5.
[24] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 102–111, Jan. 2017.
[25] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1598–1607, 2020.
[26] H. Wang et al., "TeCANet: Temporal-contextual attention network for environment-aware speech dereverberation," 2021, *arXiv:2103.16849*.
[27] R. Zhou, W. Zhu, and X. Li, "Speech dereverberation with a reverberation time shortening target," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.
[28] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing," in *Proc. Speech Commun. 13th ITG-Symp.*, 2018, pp. 1–5.
[29] A. Li, C. Zheng, C. Fan, R. Peng, and X. Li, "A recursive network with dynamic attention for monaural speech enhancement," in *Proc. Interspeech*, 2020, pp. 2422–2426.
[30] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Proc. Interspeech*, 2022, pp. 2928–2932.
[31] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2023, pp. 2351–2364.
[32] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 5, pp. 967–977, May 2016.
[33] X. Xiang, X. Zhang, and H. Chen, "A convolutional network with multi-scale and attention mechanisms for end-to-end single-channel speech enhancement," *IEEE Signal Process. Lett.*, vol. 28, pp. 1455–1459, 2021.
[34] C.-F. R. Chen, Q. Fan, and R. Panda, "CrossVIT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 347–356.
[35] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 257–265.
[36] Y. Gou, B. Li, Z. Liu, S. Yang, and X. Peng, "CLEARER: Multi-scale neural architecture search for image restoration," *in Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17129–17140,.
[37] S. W. Zamir et al., "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14816–14826.
[38] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Mar. 2016.
[39] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.
[40] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, "Cross-scale feature fusion for object detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 431–435, Mar. 2021.
[41] W. Huang, G. Li, Q. Chen, M. Ju, and J. Qu, "CF2PN: A cross-scale feature fusion pyramid network based remote sensing target detection," *Remote Sens.*, vol. 13, no. 5, 2021, pp. 847–868.
[42] M. Suin, K. Purohit, and A. N. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3603–3612.
[43] 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.
[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[45] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard
[46] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.
[47] E. A. P. Habets, "Room impulse response generator," Technische Univ. Eindhoven, Eindhoven,The Netherlands, Tech. Rep., vol. 2, no. 2.4, p. 1, 2006.
[48] S. Guan et al., "Libri-adhoc40: A dataset collected from synchronized ad-hoc microphone arrays," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021, pp. 1116–1120.
[49] C. Richey et al., "Voices obscured in complex environmental settings (VOICES) corpus," in *Proc. Interspeech*, 2018, pp. 1566–1570.
[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[51] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.
[52] 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.
[53] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Rec. ITU-T P.862, 2001.
[54] 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.
[55] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH*, 2020, pp. 5036–5040.
[56] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[57] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1240–1253, Dec. 2017.

Lei Zhao received the B.Eng. and M.Eng. degrees in electronic & information engineering and information & communication engineering from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2019 and 2022, respectively. He is currently working toward the Ph.D. degree in information and communication engineering with Northwestern Polytechnical University, Xi'an, China. His master's research focuses on signal processing in wireless communications, especially on millimeter-wave beamforming and hybrid precoding. His current research interests include pattern recognition algorithms and speech signal processing.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.

Wenbo Zhu received the bachelor's and master's degrees from Northwestern Polytechnical University, Xi'an, China, in 2019 and 2022. His research interests include speech enhancement, speech dereverberation, and acoustic signal processing.

Shengqiang Li received the master's degree in signal and information processing from Northwestern Polytechnical University, Xi'an, China, in 2022. His research interests include speech recognition and speech translation.

Hong Luo received the master's degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2002. She is currently the General Manager of Home Screen-related Product Department, China Mobile (Hangzhou) Information Technology Company, Ltd. Her research interests include mobile communication services, digital home, and artificial intelligence. She is active in global and domestic standardization work, with 28 authorized patents. She is a Member of the Chinese Institute of Electronics and China Institute of Communications.

Xiao-Lei Zhang (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Tsinghua University, Beijing, China, in 2012. He was a Postdoctoral Researcher with Perception and Neurodynamics Laboratory, The Ohio State University, Columbus, OH, USA. He is currently a Full Professor with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China. His research interests include speech processing, underwater acoustic signal processing, machine learning, statistical signal processing, and artificial intelligence. He is a Member of IEEE SPS and ISCA.

Susanto Rahardja (Fellow, IEEE) is currently a Professor of engineering cluster with the Singapore Institute of Technology, Singapore, and Ph.D. Advisor with Northwestern Polytechnical University, Xi'an, China. His research interests include multimedia coding and processing, wireless communications, discrete transforms, machine learning, signal processing algorithms, and implementation and optimization. He contributed to the development of a series of audio compression technologies, such as Audio Video Standards AVS-L, AVS-2, ISO/IEC 14496-3:2005/Amd.2:2006, and ISO/IEC 14496-3:2005/Amd.3:2006, which have licensed worldwide. Mr. Rahardja has more than 15 years of experience in leading a research team in the above-mentioned areas. He was an Associate Editor for IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, and Senior Editor for IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He is an Associate Editor for Elsevier *Journal of Visual Communication and Image Representation*, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE TRANSACTIONS ON MULTIMEDIA, and Member of Editorial Board of IEEE ACCESS. He is a Fellow of the Academy Engineering, Singapore.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on April 28,2024 at 02:11:35 UTC from IEEE Xplore. Restrictions apply.