# Monaural Speech Dereverberation Using Temporal Convolutional Networks with Self Attention

Yan Zhao, *Student Member, IEEE,* DeLiang Wang, *Fellow, IEEE,* Buye Xu, *Member, IEEE,* and Tao Zhang, *Senior Member, IEEE*

***Abstract*—In daily listening environments, human speech is often degraded by room reverberation, especially under highly reverberant conditions. Such degradation poses a challenge for many speech processing systems, where the performance becomes much worse than in anechoic environments. To combat the effect of reverberation, we propose a monaural (single-channel) speech dereverberation algorithm using temporal convolutional networks with self attention. Specifically, the proposed system includes a self-attention module to produce dynamic representations given input features, a temporal convolutional network to learn a nonlinear mapping from such representations to the magnitude spectrum of anechoic speech, and a one-dimensional (1-D) convolution module to smooth the enhanced magnitude among adjacent frames. Systematic evaluations demonstrate that the proposed algorithm improves objective metrics of speech quality in a wide range of reverberant conditions. In addition, it generalizes well to untrained reverberation times, room sizes, measured room impulse responses, real-world recorded noisy-reverberant speech, and different speakers.**

***Index Terms*—Dereverberation, temporal convolutional networks, self attention, room impulse response.**

## I. Introduction

In real-world environments, when people converse in a room or communicate with a device, the speech signal is inevitably distorted by its delayed and damped reflections from various surfaces (walls, ceilings, tables, and so forth) during sound propagation. This type of distortion, namely, room reverberation, degrades the speech quality and intelligibility for human listeners [19], especially when the reverberation time ($T_{60}$) is long. Moreover, reverberation poses a serious problem for many speech-related applications including automatic speech recognition (ASR), which is widely utilized in smart speakers and in-car systems for voice control. It has been shown that the performance of ASR systems is severelly degraded under far-field conditions [49]. Therefore, reducing the effect of reverberation is beneficial for both human listeners and machine perception systems. In this study, we address room reverberation in monaural scenarios, which is easier to apply but more challenging lacking spatial information provided by a microphone array.

Due to the importance of speech dereverberation, it has been extensively studied and many algorithms have been developed in the past decades. For example, by assuming an exponentially decaying model of room impulse response (RIR), Lebart *et al.* [25] proposed a power spectral subtraction algorithm to remove late reverberation. Wu and Wang [47] proposed a two-stage dereverberation algorithm, which employed an inverse filter to reduce early reflections in the first stage and spectral subtraction to remove long-term reverberation in the second stage. In [28], López *et al*. considered the magnitude spectrum of late reverberation as a sparse linear combination of the magnitude spectrum of past frames and proposed to predict the late reverberation using a Lasso based approach. In order to obtain an inverse filter in short-time Fourier transform (STFT) domain, Nakatani *et al.* [32] employed a long-term linear prediction method, and achieved good dereverberation performance [6], [24]. Specifically, based on a relatively large number of past frames, frequency-dependent linear prediction filters were first estimated by minimizing the weighted prediction error (WPE). Then the enhanced signal was obtained by subtracting the filtered signal (the estimated late reverberation) from the reverberant signal. Since it operates in the complex domain, both the magnitude and phase information could be recovered. Although few distortions were introduced in WPE-processed signals, a certain amount of reverberation remains.

In recent years, deep neural networks (DNNs) have been widely used in speech enhancement or separation, and substantially outperformed conventional enhancement methods [44]. For speech dereverberation, Han *et al*. [13], [14] first proposed to learn a spectral mapping from the log magnitude spectrum of reverberant speech to that of anechoic speech by using a DNN. Noting the importance of reverberation dependent parameters in supervised training, Wu *et al*. [46] developed a reverberation-time-aware approach to suppress reverberation in a wide range of reverberation times. Better performance over Han *et al*.'s system was reported. However, the utilization of feed-forward DNNs in these approaches can only capture limited contextual information. To overcome such a limitation, Santos and Falk [39] proposed a recurrent neural network (RNN) to capture long-term contexts while employing a 2-D convolutional layer to extract short-term contextual information. Their study reported some benefits from residual connections. By employing RNN with long short-term memory (LSTM), we previously estimated the mag-

Manuscript received xxx; revised xxx.

Y. Zhao is with the Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH, 43210 USA. e-mail: zhao.836@osu.edu.

D. L. Wang is with the Department of Computer Science and Engineering and the Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH, 43210 USA. He also held a visiting appointment at the Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Xi'an, China. e-mail: dwang@cse.ohio-state.edu.

B. Xu was with Starkey Hearing Technologies, Eden Prairie, MN 55344 USA. He is now with Facebook Reality Labs, Facebook, Inc., Redmond, WA 98052 USA. e-mail: xub@fb.com.

T. Zhang is with Starkey Hearing Technologies, Eden Prairie, MN 55344 USA. e-mail: Tzhang28@ieee.org.

nitude spectrum of late reverberation first, and then subtracted it from the magnitude spectrum of reverberant speech [52]. Our training used the magnitude spectrum of direct sound plus early reflections as the training target. It is worth noting that this design can be viewed as implicitly adding skip connections from LSTM's input layer to the output layer.

Despite the advantage of incorporating long range contexts, one drawback of RNN-based approaches is that the output of the current time step depends on the computation of the previous time steps, which prevents parallelization. On the other hand, stacking multiple layers in convolutional neural networks (CNNs) also captures contextual information. Another benefit is that the hierarchical structure of CNN enables a short path to incorporate distant information than the chain structure of RNN [9]. Systematic evaluations [2] show better performance of convolutional architecture for sequence modelling. This motivates us to develop a CNN-based system for speech dereverberation.

As indicated in a previous study [46], to deal with various reverberant environments, some reverberation time dependent parameters should be included. Instead of using a reverberation time estimator, we believe that such information can be encoded in the relationship among input features (e.g., the magnitude spectrum) extracted from reverberant speech. This inspires us to apply an attention mechanism [43] to input features. By exploring the relevance among features at different time steps, the attention mechanism is expected to produce a dynamic representation according to different reverberant environments, e.g., different $T_{60}$s, and direct-to-reverberation ratios (DRRs). We take an attention layer as a feature enhancement module for the rest of the dereverberation system.

Based on the above analyses, we propose a monuaral speech dereverberation algorithm using temporal convolutional networks (TCNs) with self attention. More specifically, a self-attention module is first applied to raw input features to generate dynamic representations, and then a temporal convolutional network is employed to learn the nonlinear mapping from the enhanced features to the magnitude spectrum of anechoic speech. Finally, a 1-D convolutional layer is added to smooth the estimated magnitude spectrum among adjacent frames. A recent study [36] employing wide residual networks is related to our algorithm, but we use a different network architecture and introduce the self-attention mechanism. It is worth noting that TCN has been successfully used for speaker separation [29] and speech enhancement [33], both in the time domain.

The rest of the paper is organized as follows. We first describe our proposed algorithm in Section II. Then the experimental setup is introduced in Section III. In Section IV, the evaluation results and comparisons are presented. We conclude this paper in Section V.

## II. Algorithm Description

Fig. 1 shows the diagram of the proposed system. In this section, we first describe the signal model of reverberant speech and feature extraction. Then the details of each component of the system are introduced in the following subsections.

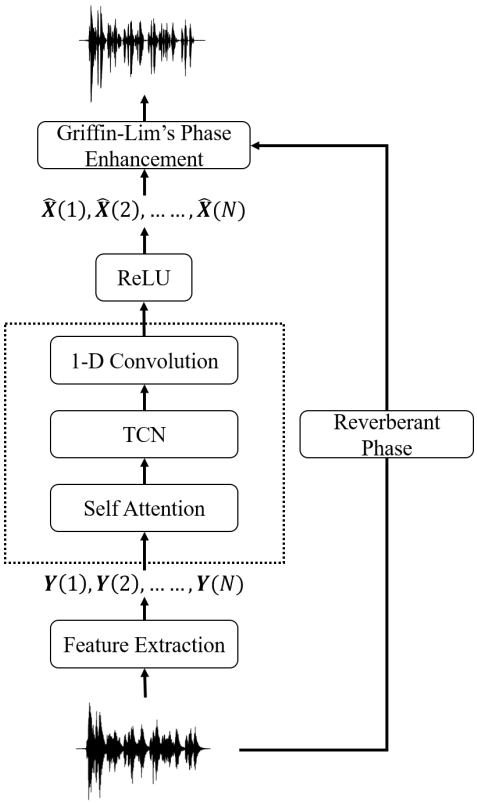



Fig. 1: Diagram of the proposed system, where $\boldsymbol{Y}(i)$ denotes magnitude spectrum and $\hat{\boldsymbol{X}}(i)$ enhanced magnitude spectrum. The trainable neural network is shown within the dotted box.

### A. Signal Model and Feature Extraction

Let $s(t)$, $y(t)$, and $h(t)$ denote clean speech, reverberant speech, and room impulse response function, respectively. The reverberant speech $y(t)$ can be written as

$$y(t) = s(t) * h(t) \tag{1}$$

where $*$ stands for the convolution operator. Here we divide $h(t)$ into two components, namely, impulse response function $h_d(t)$ for direct sound and $h_r(t)$ for reverberation. Therefore, the reverberant signal can be represented by

$$y(t) = s(t) * h_d(t) + s(t) * h_r(t) = x(t) + r(t) \tag{2}$$

Our study aims to recover the anechoic signal (direct sound) $x(t)$ given the corresponding reverberant observation $y(t)$. Note that $x(t)$ is slightly different from $s(t)$ by a time shift and an energy decay caused by sound propagation through the direct path.

Given a time-domain signal sampled at 16 kHz, we divide it into frames by using a 32 ms Hamming window with 8 ms window shift. A 512-point fast Fourier transform (FFT) is applied to each frame, which results in 257 frequency bins. To compress the dynamic range, the cubic-root compressed magnitude spectrum of the reverberant speech is used as features. We use $\boldsymbol{Y}(m)$ to denote the compressed magnitude spectrum features at time frame $m$, which is a 257-D vector. Then, our system takes the following consecutive feature

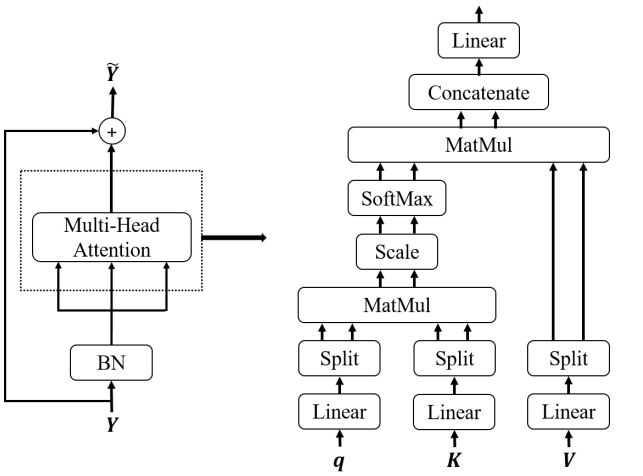



Fig. 2: Diagram of the self-attention module, where $\boldsymbol{Y}$ denotes magnitude spectrum features and $\tilde{\boldsymbol{Y}}$ enhanced features; BN denotes batch normalization. The multi-head attention module is further illustrated on the right, where the number of heads is 2.

vectors as the input,

$$\boldsymbol{Y} = \{\boldsymbol{Y}(1), \boldsymbol{Y}(2), ..., \boldsymbol{Y}(N)\} \tag{3}$$

where $N$ is the total number of frames in an utterance.

We use the cubic-root compressed magnitude spectrum of the anechoic speech as the training target. At frame $m$, let $\boldsymbol{X}(m)$ denote the compressed magnitude spectrum of the anechoic speech. The training target is expressed as

$$\boldsymbol{X} = \{\boldsymbol{X}(1), \boldsymbol{X}(2), ..., \boldsymbol{X}(N)\} \tag{4}$$

The dereverberation task is now formulated to be a sequence-to-sequence mapping problem, i.e., $\{\boldsymbol{Y}(i)\} \rightarrow \{\boldsymbol{X}(i)\}, i = 1, 2, ..., N$.

### *B. Self Attention as a Pre-processing Module*

Physically, reverberation consists of a collection of sound reflections from surfaces in an acoustic space. In other words, at each time step $t$, the sound reaching the receiver (microphone or ear) includes the attenuated and delayed past signals in addition to the direct sound, resulting in strong feature correlations at different time steps. Reverberation information like reverberation time is embedded in the correlations among different time steps of the input sequence $\{\boldsymbol{Y}(i)\}$. Instead of selecting specific models based on parameter estimation, we propose to learn a dynamic representation by exploring the input sequence to adapt to various reverberant environments. A natural choice is to introduce the attention mechanism, which dynamically emphasizes more relevant features.

Recently, attention-based models have been successfully utilized in many applications and achieved impressive performance, including machine translation [43], language understanding [7], [41], speech recognition [4], [42], speech enhancement [15], [26]. We adopt the multi-head attention mechanism [43] as a self-attention module. Fig. 2 shows the diagram of the module. Raw sequence features $\boldsymbol{Y}$ are taken as the input, then passed through a batch normalization layer (BN). After feeding the normalized features to the multi-head attention module and residual connections, a new sequence of representations $\tilde{\boldsymbol{Y}}$ is generated as the output.

Inside the multi-head attention module, query ($\boldsymbol{q}$) and key-value ($\boldsymbol{K} - \boldsymbol{V}$) pairs provide the input. They are first linearly projected to $\boldsymbol{q'}$, $\boldsymbol{K'}$, and $\boldsymbol{V'}$, respectively. Then they are split to multiple heads to capture information in different subspaces. We denote the query, keys, and values for each head as $\boldsymbol{q'_h}$, $\boldsymbol{K'_h}$, and $\boldsymbol{V'_h}$, respectively, where $h = 1, 2, .., M$, and $M$ is the number of heads. For a given query $\boldsymbol{q'_h}$, a weight distribution on the whole sequence is computed based on the similarities between the query ($\boldsymbol{q'_h}$) and keys ($\boldsymbol{K'_h}$). For each query, a compact dynamic representation is learned to incorporate more relevant information in the sequence by making a weighted sum of the values ($\boldsymbol{V'_h}$). We use the scaled dot product to measure the similarities. Therefore, the attention is computed by

$$Attention(\boldsymbol{q'_h}, \boldsymbol{K'_h}, \boldsymbol{V'_h}) = SoftMax(\frac{\boldsymbol{q'_h}\boldsymbol{K'_h}}{\sqrt{d_{\boldsymbol{k'_h}}}})\boldsymbol{V'_h} \tag{5}$$

where $d_{\boldsymbol{k'_h}}$ is the dimension of a vector in matrix $\boldsymbol{K'_h}$. The matrix form is written as follows,

$$Attention(\boldsymbol{Q'_h}, \boldsymbol{K'_h}, \boldsymbol{V'_h}) = SoftMax(\frac{\boldsymbol{Q'_h}\boldsymbol{K'_h}}{\sqrt{d_{\boldsymbol{k'_h}}}})\boldsymbol{V'_h} \tag{6}$$

Then, we merge the multi-head attention by concatenating these attention vectors. Finally, we obtain the new representations by passing the merged attention to a linear layer. Note that all the operations are applied to the same sequence (i.e., normalized $\{\boldsymbol{Y}(i)\}$, and $\boldsymbol{Q} = \boldsymbol{K} = \boldsymbol{V} = normalized\ \boldsymbol{Y}$). Thus we call it multi-head self-attention mechanism. We refer the reader to [43] for more details.

In our experiments, the normalized features (257-D) are first linearly projected to 256-D vectors. The number of heads is set to be 4, so the scale factor $d_{\boldsymbol{k'_h}}$ is $256/4 = 64$. After merging the 4 heads, we employ a 256-D to 257-D linear projection to recover the dimensionality. After adding the normalized $\boldsymbol{Y}$, we obtain the enhanced feature $\tilde{\boldsymbol{Y}}$ as shown in Fig. 2.

### *C. Temporal Convolutional Network as a Nonlinear Mapping Module*

We employ a temporal convolutional network (TCN) to learn the nonlinear mapping from the dynamic representation to the compressed magnitude spectrum of anechoic speech. It is worth noting that the term TCN in this paper has a slightly different meaning from that used in [2], where it is constrained to a causal setting using 1-D causal convolutions. We do not limit ourselves to such setting, i.e., 1-D non-causal convolution (normal 1-D convolution) is used to build our system. Fig. 3 shows the diagram of the TCN module and also the design of the TCN building block.

The TCN can be viewed as a deep residual network [17] using 1-D convolutional layers. The design of a TCN building block is referred to as a pre-activation design [18], where the activation function is placed before the weight layer as shown

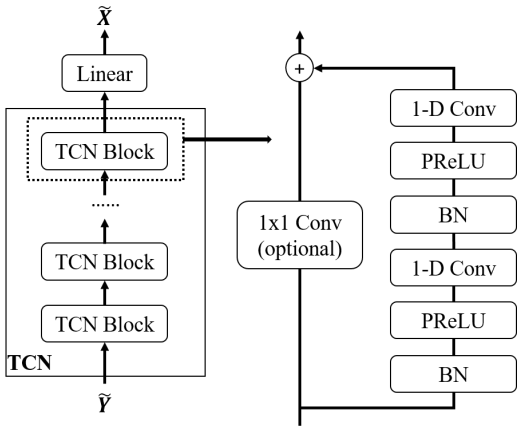



Fig. 3: Diagram of the TCN module, and a TCN block is shown on the right.

in Fig. 3. Here, Parametric Rectified Linear Unit (PReLU) [16] is utilized as the activation function. Within each block, we stack two 1-D convolutional layers. Instead of using a standard convolution, we employ depthwise separable convolution [5] in the design. By factorizing the convolution in one depthwise (channelwise) convolution and one 1x1 pointwise convolution, depthwise separable convolution dramatically reduces computation cost while preserving the learning capacity [5], [20]. The 1x1 convolutional layer in the residual connections is used only when the number of channels of the TCN block input is different from that of output. One linear layer is placed on top of the TCN to map the output of the TCN to the magnitude spectrum.

Due to the smearing effect in magnitude spectrum caused by reverberation, especially when the reverberation time is long, it is important to incorporate a large context when learning the mapping function to perform dereverberation. Although some global information of the sequence has been extracted in the self-attention module, it is still useful to enlarge the receptive field of the TCN. For CNN-based models, stacking more convolutional layers to make the network deeper is a straightforward way to increase the size of the receptive field. However, very deep architecture would cause difficulties for optimization and also make the model size large. An alternative way is to introduce dilated convolution [50] which can expand the receptive field exponentially. With these considerations, we add dilation in 1-D convolutional layers with repeating dilation rates 1, 2, 5, 9 (these numbers have no common divisor). This set of dilation rates is selected to avoid gridding artifacts caused by stacking dilated convolutional layers [45].

In our experiments, we stack 8 blocks to build the TCN module. In other words, we have 16 ($8\times2$) convolution layers in the TCN module, so the chosen dilation rates are repeated 4 times in sequential order. For each convolution layer, the kernel size is 3, the stride is 1, and the number of channels is set to 512. Therefore the top linear layer performs linear projection from 512-D to 257-D. We use the same padding for each convolution layer to keep the length of the feature sequence the same as that of the input sequence. The size of the receptive field of the TCN module is 137.

### D. 1-D Convolution as a Smoothing Module

Due to overlaps between the adjacent frames, the enhanced magnitude spectrum should be smooth. To improve the smoothness, we add one 1-D convolutional layer (depthwise separable convolution) with a small kernel size (3 in our experiments) on top of the TCN module, which can be viewed as mapping from a small context window of magnitude spectrum to the central one. Since the values of magnitude spectrum are positive, Rectified Linear Unit (ReLU) [10] is used as the activated function for the output layer.

The mean squared error (MSE) is used as the loss function, namely,

$$\mathcal{L}(\boldsymbol{Y}, \boldsymbol{X}; \Theta) = \|\boldsymbol{X} - f(\boldsymbol{Y})\|_2^2 \tag{7}$$

where $\|\bullet\|_2$ denotes the $l_2$ norm and $\Theta$ denotes the parameters in the model $f$, which are learned during training.

### E. Time-domain Signal Resynthesis

After obtaining the enhanced magnitude spectrum $\hat{\boldsymbol{X}}$, in order to reduce the STFT inconsistency with the reverberant phase, we employ Griffin-Lim's iterative phase enhancement algorithm [11] and overlap-add (OLA) method to resynthesize the corressponding enhanced time-domain signal [14]. With the better enhanced spectral magnitude, our approach is expected to deliver better enhancement performance.

### F. Causal Setting

Although the proposed dereverberation algorithm is non-causal, it can be converted to a causal system with the following modifications.

1) Applying a masking matrix before the SoftMax layer to eliminate future time steps. The values of the matrix in the upper triangle are set to be $-\infty$, and otherwise 0. Adding this masking matrix turns the future values after any time step to negative infinity while keeping the past and current values unchanged. Through softmax normalization, the weights on the future frames will be 0, making the self-attention module causal. This technique is used in the decoder part of the Transformer model [43].

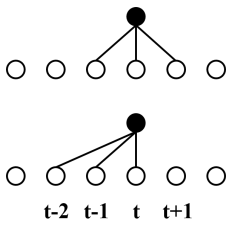



Fig. 4: Non-causal 1-D convolution (up) and causal 1-D convolution (down).

2) Replacing all the 1-D convolutions by 1-D causal convolutions. Fig. 4 shows the difference between these two types of convolution. After switching to causal convolutions, the TCN module and the smoothing module both become causal. Note

that the batch normalization requires the whole sequence to compute mean and standard deviation statistics, which is not causal during the training stage. However, during the test stage, these computed statistics are directly employed, i.e. there is no need to consider the whole sequence.

With the above modifications, the magnitude spectrum enhancement part of the proposed system becomes causal. However, Griffin-Lim's phase enhancement algorithm requires adjacent frames including several future frames, which prevents the whole system from being deployed in real time. This can be avoided by using causal phase reconstruction methods [3] or simply the reverberant phase.

## III. Experimental Setup

### A. Datasets

We evaluate our proposed system using the WSJ0 corpus [35]. The configuration for dataset generation is listed in Table I. Specifically, there are 7138, 410 and 330 clean utterances to generate training data, validation data and test data, respectively. For speakers, the training data consists of 83 speakers; the validation data consists of 10 speakers; and the test data consists of 12 speakers. Therefore, a speaker-independent dereverberation system is developed. We simulate three reverberant rooms of different sizes, from small to large, to generate the reverberant speech. The microphone is placed in a fixed position in each room while the position of the speaker is randomly chosen. As for the distance between the microphone and the speaker, we consider two situations, i.e., near (0.5 m) and far (2 m). In addition, a wide range of reverberation times is studied, from 0.3 s to 1.0 s, with a 0.1 s increment. In summary, there are 3 (room sizes) $\times$ 2 (distances) $\times$ 8 ($T_{60}$s) = 48 combinations in total. In training/validation data, for each utterance, we randomly select 6 combinations to generate 6 different RIRs, and then convolve with the clean speech to produce 6 kinds of reverberant speech. Therefore, the training set includes 7138 (clean speech) $\times$ 6 (RIRs) = 42828 utterances, and the validation set 410 (clean speech) $\times$ 6 (RIRs) = 2460 utterances. In order to investigate the gernalization ability of the proposed system, different test sets are employed (see details in Section IV). It should be pointed out that both the RIRs and sentences used for testing are unseen during training/validation. All simulated RIRs are generated utilizing an RIR generator [12], which is based on the image method [1].

We also train and evaluate the proposed system on the REVERB challenge data [24] for the monuaral speech enhancement track. The challenge data is based on WSJCAM0 corpus [38] and MC-WSJ-AV corpus [27]. Specifically, a multi-condition training set is generated by convolving 7861 clean utterances from WSJCAM0 with 24 measured RIRs. The range of reverberation times for training is from around 0.2 s to 0.8 s. Different from our simulated dataset, a moderate level of noise is included in the training set. Recorded noise is added to the reverberant speech at 20 dB signal-to-noise ratio (SNR). For development and evaluation sets, simulated data (SimData) and real recorded data (RealData) are provided. SimData includes 3 (room sizes) $\times$ 2 (distances) = 6 reverberant conditions. The reverberation times of these three rooms are about 0.3 s, 0.6 s and 0.7 s. In the development set, 1484 sentences from the WSJCAM0 corpus are convolved with 6 RIRs to generate reverberant speech and background noise is also added at SNR of 20 dB. The simulated evaluation data is generated in a similar way with 2176 sentences from the WSJCAM0 corpus. RealData consists of real recordings in a noisy and reverberant room. The reverberation time is about 0.7 s. There are 1 (room size) $\times$ 2 (distances) = 2 reverberant conditions. Different rooms are used for the training set, development set, and evaluation set, ensuring that the reverberant conditions for evaluation are unseen during system development. The background noise used in the data generation was recorded in a real room and mainly generated by air conditioning.

Our experiments focus on studying monaural speech dereverberation. Therefore, we consider each channel signal in the the multi-conditional training set as an independent reverberant signal, and utilize all the eight channels data for training. the one-channel SimData of the development set is used as the validation data during training. The trained model is evaluated using the one-channel SimData and RealData in the evaluation set.

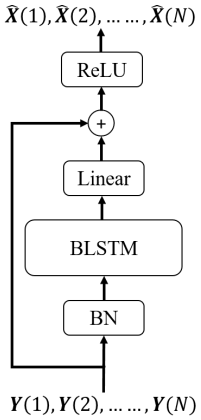



Fig. 5: Diagram of the BLSTM dereverberation baseline system.

### B. Comparison Systems

We employ an improved model of our previous study [52] as a baseline system. Fig. 5 shows the diagram of the system. It is denoted by "**BLSTM**" for convenience. Compared with the original method in [52], there are a couple of improvements. Firstly we replace the uni-directional LSTM with a bi-directional LSTM (BLSTM). Secondly, as we mentioned earlier, instead of explicitly performing spectral subtraction in the network, we utilize a residual learning, which implicitly forces the BLSTM to estimate the reverberation and remove it from reverberant magnitude spectrum features. Moreover, both late reverberation and early reflections are removed. In our experiments, a two layer BLSTM network with 1024 hidden units is used, with 512 units assigned to each direction. The number of the parameters of the baseline model is around 9.72M. We denote the proposed system by "**proposed**", which

TABLE I
CONFIGURATION FOR SIMULATED DATA GENERATION.

| item | configuration |
|---|---|
| clean utterance (WSJ0) | train: 7138 / validation: 410 / test: 330 |
| speakers | train: 83 / validation: 10 / test: 12 |
| room size (m×m×m) | 5.6×3.8×2.5 (small) / 6.3×4.9×2.6 (medium) / 6.2×6.7×3.0 (large) |
| microphone-speaker distance (m) | 0.5 (near) / 2 (far) |
| $T_{60}$ (s) (training/validation) | 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 |

TABLE II
APPROXIMATE DRRS AT DIFFERENT REVERBERATION TIMES.

| $T_{60}$ (s) | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|
| Test A (dB) | -0.38 | -2.31 | -4.61 | -4.54 | -5.40 | -6.34 | -7.01 | -7.58 |

contains around 4.70M parameters. Therefore, from the perspective of model size, the proposed model is substantially smaller than the baseline model.

In addition, an ablation study is conducted to investigate the effects of the three modules in the proposed system. More specifically, three systems using the nonlinear mapping module, the mapping module plus the smoothing module, the mapping module plus the self-attention module are included for comparisons and denoted by "**TCN**", "**TCN+smoothing**", and "**TCN+self-attention**", respectively.

### C. Optimization

The systems are initialized by using the orthogonal initialization method [40], and trained using the Adam [23] optimizer with a weight decay. For the proposed CNN-based system, twelve utterances are batched together, where the shorter utterances are padded by repeating themselves. Performing this padding provides a better estimate of statistics for batch normalization than zero padding. The padded parts are removed at the output when computing the loss. For the baseline BLSTM system, similar to the previous study [52], a 0.3 dropout rate between the two hidden layers of the BLSTM [51] is applied to mitigate the overfitting issue. We employ the weight-dropped technique [31] with 0.5 dropout rate to mitigate the overfitting across the recurrent connections. Sixteen utterances make up one batch. Instead of zero/repeating padding to the same length or truncating the utterances with fixed length, we perform batch processing using variable length sequences supported by PyTorch [34].

### D. Evaluation Metrics

We use perceptual evaluation of speech quality (PESQ) [37] and frequency-weighted segmental signal-to-noise ratio ($\text{SNR}_{fw}$) [30] as the main metrics to evaluate the proposed approach. For both of these standard metrics, a higher value indicates better performance. For PESQ evaluation, we employ wide-band PESQ [21] as used in [6] instead of the more popularly used narrow-band version (see e.g. [44]). When evaluating a corrupted or enhanced signal, wide-band PESQ typically produces a lower value than the narrow-band counterpart.

In order to compare with other published results on the REVERB challenge data, we also report cepstrum distance (CD) [30], log likelihood ratio (LLR) [30], and speech-to-reverberation modulation energy ratio (SRMR) [8] on that dataset. For the first two metrics, the lower the better; for the last one, the higher the better. For RealData in the evaluation set, due to the absence of reference signals, only SRMR results are reported.

## IV. EVALUATIONS AND COMPARISONS

In this section, we evaluate our proposed system with both our simulated data and sim/real data provided by REVERB challenge. The system performance is compared with the baseline BLSTM, as well as TCN, TCN+smoothing, and TCN+self-attention described in Sect. III.B.

### A. Evaluations at Different Reverberation Times

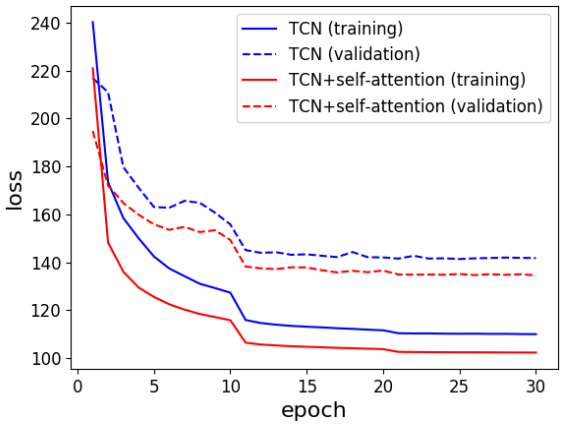



Fig. 6: Loss curves with respect to epoches for training and validation.

To test whether the trained model generalizes well to other rooms, a set of RIRs is generated for testing. We simulate a room of size 10 m × 7 m × 3 m. Eight reverberation times from 0.3 s to 1.0 s are considered, and they match the reverberation times used for training. A 2 m microphone-speaker distance is adopted to represent the far-field condition, which is more difficult than the near-field condition. Microphone position is fixed while the position of the speaker is randomly chosen. We denote this test set as "Test A". Table II shows the DRR values at different reverberation times. The DRR values are approximated on the basis of simulated RIRs according to

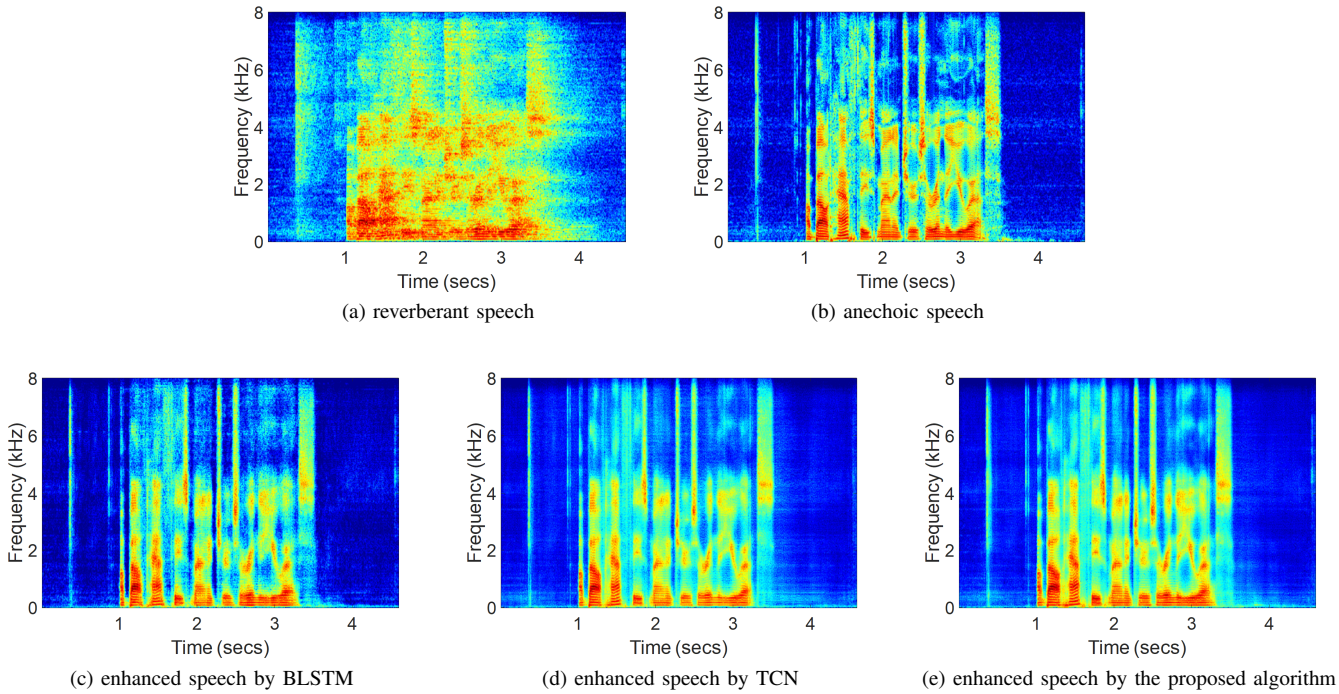
$$DRR = 10log_{10}(\frac{\sum h_d^2}{\sum h_r^2}) \quad (8)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2020.2995273, IEEE/ACM Transactions on Audio, Speech, and Language Processing



TABLE III
Average PESQ and $\text{SNR}_{fw}$ scores on Test A. Boldface number indicates the best performance.

| | **PESQ** | | | | | | | | | **$\text{SNR}_{fw}$ (dB)** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $T_{60}$ (s) | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 | **Avg.** | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 | **Avg.** |
| **reverb** | 1.75 | 1.49 | 1.29 | 1.28 | 1.24 | 1.21 | 1.21 | 1.18 | 1.33 | 10.79 | 9.07 | 6.18 | 5.68 | 4.86 | 3.88 | 2.97 | 2.54 | 5.75 |
| **BLSTM** | 2.76 | 2.58 | 2.24 | 2.11 | 2.01 | 1.93 | 1.84 | 1.78 | 2.16 | 14.98 | 13.50 | 12.89 | 12.17 | 11.58 | 10.76 | 10.33 | 10.19 | 12.05 |
| **TCN** | 2.93 | 2.82 | 2.48 | 2.39 | 2.28 | 2.19 | 2.09 | 2.01 | 2.40 | 14.23 | 13.23 | 12.18 | 11.23 | 10.64 | 9.80 | 9.03 | 8.97 | 11.16 |
| **TCN+smoothing** | 2.90 | 2.78 | 2.46 | 2.37 | 2.24 | 2.17 | 2.05 | 1.99 | 2.37 | 14.36 | 13.24 | 12.02 | 11.06 | 10.35 | 9.59 | 8.73 | 8.73 | 11.01 |
| **TCN+self-attention** | 3.02 | 2.89 | 2.53 | 2.46 | 2.31 | 2.25 | 2.14 | 2.05 | 2.46 | 15.10 | **14.28** | 13.01 | 12.04 | 11.42 | 10.72 | 9.88 | 9.94 | 12.05 |
| **proposed** | **3.06** | **2.96** | **2.60** | **2.52** | **2.38** | **2.29** | **2.18** | **2.09** | **2.51** | **15.36** | 14.25 | **13.25** | **12.69** | **11.99** | **11.42** | **10.83** | **10.59** | **12.55** |




(a) reverberant speech (b) anechoic speech

(c) enhanced speech by BLSTM (d) enhanced speech by TCN (e) enhanced speech by the proposed algorithm

Fig. 7: (Color online) Spectrograms of (a). a reverberant utterance ($T_{60}$ = 1.0 s), (b). an anechoic utterance, (c). an enhanced utterance processed by BLSTM, (d). an enhanced utterance processed by TCN, and (e). an enhanced utterance processed by the proposed algorithm.

DRR can be considered as another metric to measure the effects of reverberation besides reverberation time. It is inversely proportional to the square of the distance. A lower value indicates that it is more challenging to perform dereverberation.

Table III lists the average PESQ and $\text{SNR}_{fw}$ scores on the Test A set, where a wide range of reverberation times is evaluated. Compared to unprocessed reverberant conditions, all the systems show significant improvements in each condition in terms of both PESQ and $\text{SNR}_{fw}$. Our proposed system performs the best among the five. On average, compared with the reverberant speech, the enhanced speech processed by the proposed system improves PESQ by 1.18, and $\text{SNR}_{fw}$ by 6.8 dB, demonstrating its effectiveness to perform dereverberation. Although the TCN yields close PESQ scores, it is worse than the proposed system by more than 1 dB in $\text{SNR}_{fw}$. BLSTM shows better $\text{SNR}_{fw}$ scores than TCN, however, it performs the worst on PESQ.

Adding the smoothing module alone does not improve the performance of TCN. However, this module brings a performance gain for TCN together with self attention. TCN+self-attention performs better than TCN in terms of both PESQ and $\text{SNR}_{fw}$. When we compare the training procedure between TCN and TCN+self-attention, the introduction of self-attention pre-processing speeds up the training process with faster convergence and lower validation loss, indicating a more effective integration of contextual information. Fig. 6 shows the loss curves of TCN and TCN+self-attention for 30-epoch training, clearly demonstrating the efficiency of the self-attention module. For most speech processing tasks, especially speech dereverberation, the contextual information is important. In TCN, we enlarge receptive fields by stacking convolution layers and adding dilations in them. As the network becomes deeper, longer contexts are available. However, our pre-processing module employing the self-attention mechanism is able to leverage the contextual information more efficiently, since each frame can observe the entire sequence without stacking many layers. As Wu *et al.* pointed out [46], for reverberant speech the inter-frame correlation depends on $T_{60}$ and the correlation of neighboring frames becomes weaker for shorter reverberation time. We find that, after

pre-processing with the self-attention module, the inter-frame correlation is somewhat enhanced between the frames with significant anechoic speech energy. This helps to enhance the dereverberation performance.

In Fig. 7, one example is given to illustrate the effectiveness of the proposed system to perform dereverberation. We randomly select one utterance from Test A with the most severe reverberant condition ($T_{60}$ is 1.0 s and the microphone-speaker distance is 2 m). Fig. 7(a) presents the spectrogram of the reverberant speech. The corresponding spectrogram of the anechoic speech is shown in Fig. 7(b). The enhanced speech processed by BLSTM, TCN, and the proposed algorithm are shown in Fig. 7(c), (d) and (e), respectively. It is clear that most smearing effects caused by reverberation have been suppressed and the spectrotemporal structure of the corrupted speech is well enhanced by all three algorithms, with the proposed algorithm recovering the most structure of the anechoic utterance. This demonstrates that our approach is able to remove reverberation effects in adverse reverberant environments.

### B. Evaluations at Untrained Long Reverberation Times

In order to test the performance under untrained reverberation times, we generate a set of RIRs with reverberation times from 1.1 s to 1.5 s within the same simulated room as in Test A. Let "Test B" denote the new test set with these long reverberation times. The DRR values are presented in Table IV. Under longer reverberation times and lower DRRs, we investigate the generalization ability of our models to very reverberant environments.

TABLE IV
APPROXIMATE DRRS AT UNTRAINED LONG REVERBERATION TIMES.

| $T_{60}$ (s) | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 |
|---|---|---|---|---|---|
| Test B (dB) | -7.81 | -8.31 | -8.59 | -9.03 | -9.27 |

The PESQ and $\text{SNR}_{fw}$ scores are presented in Table V. Similar performance trends are observed to those in Test A. Although longer reverberation time conditions are not included during training, all the systems show good generalization to the severely reverberant environments. On average, BLSTM, TCN, and the proposed system improve PESQ by 0.43, 0.62, and 0.66, $\text{SNR}_{fw}$ by 7.16, 5.86, and 7.67 dB, respectively. Again, our proposed system delivers the best dereverberation performance. The comparisons among the TCN, TCN+smoothing, TCN+self-attention, and proposed system show a similar trend to that in Test A.

To evaluate the effect of the Griffin-Lim phase enhancement algorithm in our proposed system, we resynthesize the enhanced signals using the OLA method with the reverberant phase for comparison. On average, for Test A, the phase enhancement improves PESQ by 0.17 and $\text{SNR}_{fw}$ by 0.13 dB; for Test B, it improves PESQ by 0.09 and $\text{SNR}_{fw}$ by 0.22 dB. The results indicate that the phase enhancement benefits speech dereverberation in the magnitude spectrum by partly recovering the clean phase.

In addition, we investigate the performance of the proposed system and its causal counterpart. For both systems, we directly use the reverberant phase for simplicity. On average, for Test A, the causal system underperforms the non-causal system by 0.22 in PESQ and by 1.06 dB in $\text{SNR}_{fw}$; for Test B, the causal version underperforms by 0.2 in PESQ and by 1.86 dB in $\text{SNR}_{fw}$.

### C. Evaluations with Recorded RIRs

Test A and B used simulated RIRs. Now we evaluate the trained systems with recorded RIRs. Note that they are different from the later evaluations with REVERB challenge data where measured RIRs are employed to generate training data. Three RIRs are selected from the Aachen Impulse Response (AIR) database [22], and resampled at 16 kHz. They were recorded in a lecture hall, a meeting room and an office, and the corresponding reverberation time is $0.70$ s, $0.21$ s and $0.37$ s, respectively. A new test set is generated using these recorded RIRs, which is denoted by "Test C".

Table VI shows the PESQ and $\text{SNR}_{fw}$ evaluation results of Test C. Clearly, all these systems generalize well to real reverberant rooms. Substantial improvements in objective metrics are obtained over unprocessed reverberant speech. Like previous evaluations, the proposed system outperforms the other systems, including the alternatives with removed modules. Slightly different from Test A and B, TCN performs better than BLSTM in both PESQ and $\text{SNR}_{fw}$.

The Test C results are quite encouraging. They imply that the models trained with simulated RIRs generalize well to recorded RIRs, and therefore it might be sufficient to use simulated RIRs only for training supervised speech dereverberation systems. Infinite RIRs can be produced with a simulated RIR generator, in order to increase the diversity of training data. Another benefit is cost saving from having to record RIRs in real rooms.

### D. Evaluations with REVERB Challenge Data

In this subsection, we evaluate the proposed algorithm on REVERB challenge data described in Sect. III.A. As mentioned earlier, we treat the 8-ch speech as eight different 1-ch speech recordings. The evaluation results are presented in Table VII. We compare our system with a group of WPE-based systems [6] and two DNN-based systems [48], [36]. Overall, our system performs the best among the monaural systems. Surprisingly, our algorithm performs comparably to the WPE+MVDR+MMSE (8-ch) approach [6].

Although our algorithm is for dereverberation, the results on the REVERB challenge indicate its robustness to some background noise. This is because the proposed model is trained to learn a nonlinear mapping from noisy-reverberant speech to clean-anechoic speech using data with both background noise and room reverberation (see Sect. III.A). Moreover, the SNR in this dataset is relatively high.

## V. CONCLUDING REMARKS

Reverberation creates one of the major speech distortions in daily environments. Together with background noise, it

TABLE V
Average PESQ and $\text{SNR}_{fw}$ scores on Test B.

| | **PESQ** | | | | | | **$\text{SNR}_{fw}$** (dB) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $T_{60}$ (s) | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | **Avg.** | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | **Avg.** |
| **reverb** | 1.16 | 1.18 | 1.16 | 1.14 | 1.14 | 1.16 | 2.35 | 1.54 | 1.37 | 1.02 | 0.82 | 1.42 |
| **BLSTM** | 1.71 | 1.61 | 1.62 | 1.53 | 1.49 | 1.59 | 9.58 | 9.32 | 8.66 | 8.07 | 7.28 | 8.58 |
| **TCN** | 1.95 | 1.80 | 1.83 | 1.70 | 1.64 | 1.78 | 8.56 | 7.70 | 7.45 | 6.42 | 6.26 | 7.28 |
| **TCN+smoothing** | 1.93 | 1.78 | 1.80 | 1.68 | 1.62 | 1.76 | 8.35 | 7.53 | 7.19 | 6.26 | 6.13 | 7.09 |
| **TCN+self-attention** | 1.99 | 1.84 | 1.84 | 1.72 | 1.65 | 1.81 | 9.46 | 8.69 | 8.31 | 7.26 | 7.01 | 8.15 |
| **proposed** | **2.01** | **1.85** | **1.87** | **1.73** | **1.66** | **1.82** | **10.29** | **9.84** | **9.32** | **8.35** | **7.67** | **9.09** |

TABLE VI
Average PESQ and $\text{SNR}_{fw}$ scores on Test C.

| | **PESQ** | | | **$\text{SNR}_{fw}$** (dB) | | |
|---|---|---|---|---|---|---|
| **room** | lecture | meeting | office | lecture | meeting | office |
| **reverb** | 1.38 | 1.99 | 1.60 | 6.30 | 7.52 | 7.43 |
| **BLSTM** | 2.27 | 2.89 | 2.31 | 11.36 | 9.91 | 9.80 |
| **TCN** | 2.61 | 3.15 | 2.72 | 12.71 | 10.63 | 10.49 |
| **TCN+smoothing** | 2.59 | 3.14 | 2.72 | 12.75 | 10.64 | 10.49 |
| **TCN+self-attention** | 2.65 | **3.20** | 2.73 | 12.92 | 10.48 | 10.67 |
| **proposed** | **2.69** | 3.18 | **2.77** | **13.24** | **10.95** | **10.86** |

TABLE VII
Average performance of different algorithms on SimData and RealData of REVERB challenge evaluation set.

| | SimData | | | | RealData | |
|---|---|---|---|---|---|---|
| | **CD** | **SRMR** | **LLR** | **$\text{SNR}_{fw}$** (dB) | **PESQ** | **SRMR** |
| **reverb** | 3.97 | 3.68 | 0.58 | 3.62 | 1.48 | 3.18 |
| **WPE** (1-ch) | 3.74 | 4.22 | 0.52 | 4.90 | 1.72 | 3.97 |
| **WPE** (2-ch) [6] | 3.66 | 4.50 | 0.47 | 5.35 | 1.82 | 4.48 |
| **WPE** (8-ch) [6] | 3.63 | 4.64 | 0.46 | 5.48 | 1.89 | 4.55 |
| **WPE + MVDR + MMSE** (8-ch) [6] | 2.25 | 5.39 | 0.43 | 10.31 | **2.82** | **7.34** |
| **DNN** (1-ch) [48] | 2.50 | **5.77** | 0.50 | 7.55 | - | 4.36 |
| **WRN** (1-ch) [36] | 3.59 | 3.59 | 0.47 | 4.80 | - | 3.24 |
| **proposed** (1-ch) | **2.20** | 5.17 | **0.24** | **13.06** | 2.58 | 5.54 |

degrades speech intelligibility and quality. In this study, we have proposed a monaural CNN-based speech dereverberation algorithm, which includes a self-attention module to learn a dynamic representation, a TCN module to perform nonlinear mapping, and a 1-D convolution smoothing module. Systematic evaluations demonstrate that our system suppresses reverberation effectively and the trained model generalizes well to untrained conditions. Although designed for dereverberation, it seems insensitive to a moderate amount of background noise.

Future research will extend the proposed monaural algorithm to multi-channel scenarios. With spatial information, the performance of speech dereverberation is expected to be further improved. In addition, the current study aims to perform speech dereverberation in the spectral magnitude domain, leaving the phase issue to postprocessing with the Griffin-Lim algorithm. Exploring phase enhancement within the network architecture would be another promising direction for the future.

## ACKNOWLEDGMENTS

This study was supported in part by an NIH grant (R01 DC012048), a Starkey research gift, and the Ohio Supercomputer Center. The authors would like to thank Yuzhou Liu for helpful discussions.

## REFERENCES

[1] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, 1979.
[2] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.
[3] G. T. Beauregard, X. Zhu, and L. Wyse, "An efficient algorithm for real-time spectrogram inversion," in *Proc. DAFx Conference*, 2005, pp. 116–118.
[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: a neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*, 2016, pp. 4960–4964.
[5] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proc. CVPR*, 2017, pp. 1251–1258.
[6] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *REVERB Workshop*, 2014.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[8] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. on Audio, Speech, Lang. Process.*, vol. 18, pp. 1766–1774, 2010.

[9] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. ICML*, vol. 70, 2017, pp. 1243–1252.
[10] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. AISTATS*, 2011, pp. 315–323.
[11] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. on Audio, Speech, Lang. Process.*, vol. 32, pp. 236–243, 1984.
[12] E. Habets. Room impulse response generator. Available at https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator.
[13] K. Han, Y. Wang, and D. L. Wang, "Learning spectral mapping for speech dereverberation," in *Proc. ICASSP*, 2014, pp. 4628–4632.
[14] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. on Audio, Speech, Lang. Process.*, vol. 23, pp. 982–992, 2015.
[15] X. Hao, C. Shan, Y. Xu, S. Sun, and L. Xie, "An attention-based neural network approach for single channel speech enhancement," in *Proc. ICASSP*, 2019, pp. 6895–6899.
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *Proc. ICCV*, 2015, pp. 1026–1034.
[17] ——, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.
[18] ——, "Identity mappings in deep residual networks," in *Proc. ECCV*. Springer, 2016, pp. 630–645.
[19] K. S. Helfer and L. A. Wilber, "Hearing loss, aging, and speech perception in reverberation and noise," *Journal of Speech and Hearing Research*, vol. 33, pp. 149–155, 1990.
[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
[21] ITU-T Recommendation P.862.2, *Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs*, 2007.
[22] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. ICDSP*, 2009, pp. 1–5.
[23] D. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[24] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, p. 7, 2016.
[25] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acustica*, vol. 87, pp. 359–366, 2001.
[26] C.-F. Liao, Y. Tsao, X. Lu, and H. Kawai, "Incorporating symbolic sequential modeling for speech enhancement," in *Proc. INTERSPEECH*, 2019, pp. 2733–2737.
[27] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments," in *Proc. ASRU*, 2005, pp. 357–362.
[28] N. López, Y. Grenier, G. Richard, and I. Bourmeyster, "Single channel reverberation suppression based on sparse linear prediction," in *Proc. ICASSP*, 2014, pp. 5182–5186.
[29] Y. Luo and N. Mesgarani, "Conv-TasNet: surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. on Audio, Speech, Lang. Process.*, vol. 27, pp. 1256–1266, 2019.
[30] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, pp. 3387–3405, 2009.
[31] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," *arXiv preprint arXiv:1708.02182*, 2017.
[32] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. on Audio, Speech, Lang. Process.*, vol. 18, pp. 1717–1731, 2010.
[33] A. Pandey and D. L. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. on Audio, Speech, Lang. Process.*, vol. 27, pp. 1179–1188, 2019.
[34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Autodiff Workshop*, 2017.
[35] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on speech and natural language*, 1992, pp. 357–362.
[36] D. Ribas, J. Llombart, A. Miguel, and L. Vicente, "Deep speech enhancement for reverberated and noisy signals using wide residual networks," *arXiv preprint arXiv:1901.00660*, 2019.
[37] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752.
[38] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition," in *Proc. ICASSP*, vol. 1, 1995, pp. 81–84.
[39] J. F. Santos and T. H. Falk, "Speech dereverberation with context-aware recurrent neural networks," *IEEE/ACM Trans. on Audio, Speech, Lang. Process.*, vol. 26, pp. 1236–1246, 2018.
[40] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.
[41] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "DiSAN: directional self-attention network for RNN/CNN-free language understanding," in *Proc. AAAI*, 2018.
[42] M. Sperber, J. Niehues, G. Neubig, S. Stüker, and A. Waibel, "Self-attentional acoustic models," *arXiv preprint arXiv:1803.09519*, 2018.
[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.
[44] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," *IEEE/ACM Trans. on Audio, Speech, Lang. Process.*, vol. 26, pp. 1702–1726, 2018.
[45] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proc. WACV*, 2018, pp. 1451–1460.
[46] 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. on Audio, Speech, Lang. Process.*, vol. 25, pp. 102–111, 2017.
[47] M. Wu and D. L. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. on Audio, Speech, Lang. Process.*, vol. 14, pp. 774–784, 2006.
[48] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, p. 4, 2016.
[49] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 114–126, 2012.
[50] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.
[51] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.
[52] Y. Zhao, D. L. Wang, B. Xu, and T. Zhang, "Late reverberation suppression using recurrent neural networks with long short-term memory," in *Proc. ICASSP*, 2018, pp. 5434–5438.

**Yan Zhao** (S'16) received his B.E. degree in information engineering from Xi'an Jiaotong University, Xi'an, China, in 2009, his M.S. degree in electrical and computer engineering from The Ohio State University, Columbus, OH, USA, in 2014, and his second M.S. degree in computer science and engineering from The Ohio State University, Columbus, OH, USA, in 2018. He is currently pursuing the Ph.D. degree at the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA. His research interests include speech separation and machine learning.

**DeLiang Wang, Buye Xu, and Tao Zhang**, photograph and biography not provided at the time of publication.