Monaural Speech Enhancement Using Deep Multi-Branch Residual Network with 1-D Causal Dilated Convolutions

Qiquan Zhang, Aaron Nicolson, and Mingjiang Wang

Abstract—Deep learning has achieved substantial improvement on single-channel speech enhancement tasks. However, the performance of multi-layer perceptions (MLPs)-based methods is limited by its ability to capture the long-term effective history information. Long short-term memory (LSTM) model is limited by its high computational complexity and the exploding/vanish gradients problem. In this paper, we propose a novel convolution neural network (CNN) architecture, called multi-branch residual network (MB-ResNet), for monaural speech enhancement. The MB-ResNet exploits split-transform-aggregate design, which is expected to obtain strong representational power at a low computational complexity. The combination of causal dilated one dimensional CNN and residual connection significantly expands receptive fields, which allows for capturing extremely long-term temporal contextual information. Our extensive experimental investigation suggest that the MB-ResNet outperform the residual long short-term memory networks (ResLSTMs) and the CNN networks that employ residual or dense aggregations in terms of speech intelligibility and quality, while providing superior parameter efficiency. Furthermore, our experimental results demonstrate that our proposed MB-ResNet model is able to outperform multiple state-of-the-art deep learning-based speech enhancement methods in terms of five widely used objective metrics.

Index Terms—Multi-branch networks, dilated convolutions, residual connection, speech enhancement, Deep Xi.

I. INTRODUCTION

The objective of speech enhancement is to remove the background interference and improve the overall perceived quality and intelligibility of the degraded speech. Speech enhancement is an important and challenging task in the speech processing community and is applied to a wide range of speech-related applications, e.g., robust automatic speech recognition, speaker identification, mobile speech communication, and hearing aids, etc. In this study, we focus on single-channel speech enhancement [1]–[8].

Conventional speech enhancement methods are designed based on the assumption of speech and noise characteristic, including spectral subtraction algorithms [4], [5], Wiener filtering [6], [7], statistical model-based methods [8]–[11], and non-negative matrix factorization methods [12], [13]. In the last few years, deep neural network (DNN)-based monaural speech enhancement methods [3] have received a tremendous amount of attention as they have demonstrated a significant performance improvement over traditional approaches. Inspired by the concept of time-frequency (T-F) masking in computational auditory scene analysis (CASA) [14], multi-layer perceptrons (MLPs) are firstly introduced by Wang and Wang [15] to estimate the ideal binary mask (IBM) [16]. Recently, researchers have proved that ideal ratio mask (IRM) [17] based methods are able to attain better speech quality than binary mask based methods. Instead of T-F masking, Xu et al. [18] proposed to use an MLP to map the noisy speech log-power spectra (LPS) to the clean speech LPS. For spectra or mask estimation, MLPs-based methods capture temporal information utilizing a context window. However, MLPs are not able to learn the long-term dependencies inherent in noisy speech. In order to capture the long-range dependencies of noisy speech, Chen et al. [19] proposed to use a recurrent neural network (RNN) with four hidden long short-term memory (LSTM) [20] layers to estimate the ideal speech enhancement. The LSTM model has been shown to generalize to unseen speakers well and significantly outperforms MLP-based models.

While able to model the long-range dependencies of noisy speech, LSTM-based models exhibit several drawbacks. The high memory requirement and the high computational complexity of the LSTM model significantly limits its applicability. For RNNs, the another major issue is the exploding/vanish gradients problem, which causes RNNs to be difficult to train. Over the last decade, convolutional neural networks (CNNs) have gained a considerable amount of success on computer vision and image classification tasks. Recently, CNNs have made great progress in sequence modeling tasks, e.g., speech recognition [21]. The introduction of dilated causal convolutional units enables a CNN to garner an exponentially large receptive field. The dense (DenseNet) [22] and residual connections (ResNet) [23] of CNNs allow for both very deep networks and a very long effective history. The combination of very deep networks (densely and residual connected layers) and dilation causal convolutional units has allowed CNNs [24] to outperform LSTM models across a diverse range of sequence modeling tasks. Inspired by the success of CNNs, we proposed a novel multi-branch residual network (MB-ResNet) with 1-D causal dilated convolutions for speech enhancement.
enhancement. Experimental results show that the proposed MB-ResNet is able to provide more excellent performance than several advanced networks, i.e., ResLSTM, ResNet, and DenseNet. Moreover, we also find that our proposed model substantially outperforms several state-of-the-art deep learning based speech enhancement methods.

The statistic model-based speech enhancement methods heavily depend on the estimate of the a priori SNR. Recently, a residual LSTM (ResLSTM) network was proposed to estimate the a priori SNR (Deep Xi) directly from noisy speech spectral magnitude. The estimated a priori SNR can be flexibly employed in statistic model-based methods, e.g., MMSE short-time spectral amplitude (MMSE-STSA) estimator [6], Log-spectral magnitude MMSE (MMSE-LSA) estimator [9], and the square-root Wiener filter (SRWF) [7]. Here, the MB-ResNets are used within the Deep Xi framework for performance evaluation.

The remainder of this paper will be structured as follows. In Section II, we give a description of the Deep Xi speech enhancement framework. In Section III, the proposed model is described in detail. To demonstrate the superiority of our model, we conduct the experiments on two datasets (one is ours and one is publicly available dataset used in many previous works). The experimental setup and results on these two datasets are given in Section IV and Section V, respectively. In Section VI, we provide the conclusions and discussions.

II. DEEP XI SPEECH ENHANCEMENT FRAMEWORK

A. Problem Formulation

In the time-domain, the noisy speech signal, \( y[n] \), is given by
\[
y[n] = s[n] + d[n],
\]
where \( s[n] \) and \( d[n] \) denote the clean speech and uncorrelated additive noise, respectively, and \( n \) denotes the discrete-time index. The noisy speech \( y[n] \) is then analysed frame-wise using the short-time Fourier transform (STFT):
\[
Y_l[k] = S_l[k] + D_l[k],
\]
where \( Y_l[k], S_l[k], \) and \( D_l[k] \) denote the complex-valued STFT coefficients of the noisy speech, the clean speech and the noise, respectively, for time-frame index \( l \) and discrete-frequency index \( k \). In polar form, the STFT coefficient of noisy speech is expressed as \( Y_l[k] = R_l[k]e^{i\phi_l[k]} \), where \( R_l[k] \) and \( \phi_l[k] \) are the short-time spectral magnitude and phase spectrum of noisy speech, respectively. The a priori SNR \( \xi[l, k] \) and the a posteriori SNR \( \gamma[l, k] \) are defined as follows:
\[
\xi[l, k] = \frac{\lambda_s[l, k]}{\lambda_d[l, k]}, \quad \gamma[l, k] = \frac{R^2_l[k]}{\lambda_d[l, k]},
\]
where \( \lambda_s[l, k] = E\{|S_l[k]|^2\} \) and \( \lambda_d[l, k] = E\{|D_l[k]|^2\} \) denote the spectral variances of speech and noise (\( E\{\cdot\} \), the expectation operator), respectively.

For statistical model-based speech enhancement methods, the estimate of clean speech magnitude \( |\hat{S}_l[k]| \) is obtained using a gain function as \( |\hat{S}_l[k]| = G(\xi[l, k], \gamma[l, k])R_l[k]. \) The enhanced speech magnitude spectrum and noisy speech phase spectrum are fed into a synthesizer to construct the enhanced time-domain speech waveform. The gain function depends on the assumed statistical models for the speech and the noise and on the criterion that is optimized for. Based on minimum mean square error (MMSE) criterion and the Gaussian distributions for speech and noise STFT coefficients, the gain function of widely used three speech estimators, SRWF [7], MMSE-STSA [8], and MMSE-LSA [9] are obtained as
\[
G_{\text{SRWF}}(l, k) = \sqrt{\frac{\xi[l, k]}{\xi[l, k] + 1}}
\]
\[
G_{\text{MMSE-STSA}}(l, k) = \frac{\sqrt{\pi}}{2} \frac{\sqrt{v[l, k]}}{\gamma[l, k]} \exp\left(-\frac{v[l, k]}{2}\right) \cdot \left(1 + v[l, k]\right)I_0\left(\frac{v[l, k]}{2}\right) + v[l, k]I_1\left(\frac{v[l, k]}{2}\right)
\]
\[
G_{\text{MMSE-LSA}}(l, k) = \frac{\xi[l, k]}{\xi[l, k] + 1} \exp\left\{\frac{1}{2} \int_{v[l, k]}^{\infty} e^{-t} \text{d}t\right\}
\]
where \( I_0(\cdot) \) and \( I_1(\cdot) \) are the modified Bessel function of zero and first order, respectively, and \( v[l, k] = \xi[l, k] \cdot \gamma[l, k]/(\xi[l, k] + 1) \). The gain functions depend on two parameters, i.e., the a priori SNR and a posteriori SNR.

B. Mapped A Priori SNR Training Target

As the enhancement performance is predominately affected by the accuracy of the a priori SNR estimate, the training target in Deep Xi framework [7] is the mapped a priori SNR, as described in [25]. The mapped a priori SNR is a mapped version of the oracle (or instantaneous) a priori SNR. For the oracle case, the clean speech and noise of the noisy speech in (1) are known completely. This means that \( \lambda_s[l, k] \) and \( \lambda_d[l, k] \) in (3) can be replaced with the squared magnitude of the clean speech and noise spectral components, respectively.

In [25], the oracle a priori SNR (in dB), \( \xi_{\text{dB}}[l, k] = 10\log_{10}(\xi[l, k]) \), was mapped to the interval \([0, 1]\) in order to improve the rate of convergence of the used stochastic gradient descent algorithm. The cumulative distribution function (CDF) of \( \xi_{\text{dB}}[l, k] \) was used as the map. It can be seen in [25, Fig. 2] that the distribution of \( \xi_{\text{dB}} \) for a given frequency component follows a normal distribution. It was thus assumed that \( \xi_{\text{dB}}[l, k] \) is distributed normally with mean \( \mu_k \) and variance \( \sigma_k^2 \); \( \xi_{\text{dB}}[l, k] \sim N(\mu_k, \sigma_k^2) \). The map is given by
\[
\xi[l, k] = \frac{1}{2} \left[1 + \text{erf}\left(\frac{\xi_{\text{dB}}[l, k] - \mu_k}{\sigma_k \sqrt{2}}\right)\right],
\]
where \( \xi[l, k] \) is the mapped a priori SNR. Following [25], the statistics of \( \xi_{\text{dB}}[l, k] \) for each noisy speech spectral component

*The Tensorflow implementation of Deep Xi framework is available at: https://github.com/anicolson/DeepXi.
are found over a sample of the training set. During inference, $\xi[l, k]$ is found from $\hat{\xi}_{ab}[l, k]$ as follows:

$$
\hat{\xi}[l, k] = 10^{\frac{\hat{\xi}_{ab}[l, k]}{10}},
$$

where the a priori SNR estimate in dB is computed from the mapped a priori SNR estimate as follows:

$$
\hat{\xi}_{ab}[l, k] = \sigma_k \sqrt{2} \text{erf}^{-1} \left( 2 \xi[l, k] - 1 \right) + \mu_k.
$$

III. PROPOSED MODEL

A. Dilated 1-D Causal Convolutions

In the presented network architecture, the causal convolutions are employed for real-time speech enhancement task. Causal model means that there is no leakage of information from future into past. For the conventional causal convolutions, an extremely deep network or large size kernels is necessary to build a long effective history size for augmenting the ability of capturing the contextual information. However, this introduced two typical issues: the vanishing gradient problem and raising computational burden. In [26] the dilated convolutions were firstly employed to achieve multi-scale context aggregation for semantic segmentation task, and performance improvements have been achieved. The success benefits from the dilated convolutions that enable an exponentially large receptive field, while preserving the input resolution as well as computational efficiency.

![Fig. 1: An example of 1-D dilated causal convolution with the dilation rate $d = 1, 2, 4$ and kernel size $k = 3$. The receptive field is able to cover all values from the input sequence.](image)

Formally, a 1-D discrete dilated convolution operator $*d$, which convolves a sequence input $x \in \mathbb{R}^T$ with a kernel $f \in \mathbb{R}^K$, is represented as

$$
\mathbf{x} *_d f(s) = \sum_{k=0}^{K-1} f(k) \mathbf{x}(s - d \cdot k)
$$

where $d$ is the dilation rate and $K$ denotes the kernel size. As a special case, dilated convolution with dilation rate $d = 1$ is equivalent to regular convolution. Fig. 1 illustrates an example of 1-D dilated causal convolution with kernel size $k = 3$ and dilation factor $d = 1, 2, 4$. As shown in Fig. 1, the receptive field of the network grows exponentially when the dilation factor $d$ increases exponentially with the depth of the network. It enables dilated causal convolutions to capture extremely long-term temporal contextual information using deep networks.

![Fig. 2: The illustration of two commonly used residual blocks of (a) basic structure and (b) bottleneck structure, where $+$ represents the element-wise summation operation.](image)

B. Residual Connections

As well as dilated convolutions the depth of models and the kernel size are also important for capturing a long-term temporal contextual information (large receptive field). With the depth increasing, it also comes with a challenge of vanishing gradients. In [23] He et. al introduce identity shortcut connections to design a deep residual learning framework to ease the training of networks. The residual learning has been demonstrated to be a very effective way to address the vanishing gradients issue and has commonly been regarded as the default starting point for training very deep networks. In our design of the proposed model, we therefore employ residual blocks to train a deep network. Fig. 2(a) and Fig. (b) depict the 1-D basic and bottleneck residual blocks, respectively. The identity shortcut connections add their outputs to the outputs of the stacked layers so that the all information always pass through, and are considered the key factor for the ease of training deep networks.

C. Multi-Branch ResNet

MB-ResNet, as shown in Fig. 3, consists of three modules, i.e., a fully-connected input layer, $\mathbf{FC}$, $N$ multi-branch residual blocks where $n = 1, 2, 3, \ldots$, $N$ is the block index, and a fully-connected output layer, $\mathbf{O}$.  

---

\[ \text{(1)} \]

\[ \text{(2)} \]

\[ \text{(3)} \]

\[ \text{(4)} \]

\[ \text{(5)} \]

\[ \text{(6)} \]

\[ \text{(7)} \]

\[ \text{(8)} \]

\[ \text{(9)} \]

\[ \text{(10)} \]
convolutional units on each path have a kernel size of 1, whilst the second convolutional unit has a kernel size of \( k \). The second convolutional unit has a dilation rate of \( d \), providing the capability of capturing the long-term contextual information. As in [32], the dilation rate \( d \) is cycled as the block index \( n \) increases: \( d = 2^{\left( (n-1) \mod (\log_2(D)+1) \right)} \), where mod is the modulo operation, and \( D \) is the maximum dilation rate. Such stacked residual blocks support the exponential expansion of the receptive field without loss of input resolution, which allows for capturing long-term effective history.

The feature-maps produced by different branch networks are aggregated by concatenation. Then the aggregate feature-maps is processed with a 1-D convolutional unit and the output of a multi-branch residual network is produced by using a identity residual connection. This convolutional unit is also pre-activated by layer normalisation followed by the ReLU activation function [33].

3) Implementation Details: FC is a fully-connected layer unit has an output size of \( d_{\text{model}} \), where layer normalisation applied to the output of FC, followed by the ReLU activation function. As in [34], \( k \) is set to 3, and \( D \) is set to 16. On each branch, the first and the second convolutional units have an output size of \( d_f \). The output layer, O, is a fully-connected layer with sigmoidal units. The following hyperparameters were chosen for the network architecture: \( d_{\text{model}} = 256 \) and \( d_f = 16 \). MB-ResNet with size of 1.05, 1.43, and 1.66 million parameters are formed by cascading the 12, 17, and 20 multi-branch residual blocks.

IV. EXPERIMENTAL SETUP

A. Signal Processing

A square-root-Hann window function is used for spectral analysis and synthesis [35]–[37], with a frame-length of 32 ms (512 time-domain samples) and a frame-shift of 16 ms (256 time-domain samples). The a priori SNR was estimated from the 257-point single-sided noisy speech magnitude, which included both the DC frequency component and the Nyquist frequency component. The a posteriori SNR for MMSE-based speech estimators is estimated from the a priori SNR estimate: \( \gamma(l, k) = \xi(l, k) + 1 \).

B. Baseline Models

In our experiments, we compare the proposed MB-RseNet with the following network architectures (baselines) tasked with estimating the a priori SNR for MMSE-based methods to speech enhancement (Deep Xi framework):

ResLSTM: As baselines, we use three residual LSTMs (ResLSTMs) composed of 4, 5, and 6 residual blocks, and the memory cell sizes for each ResLSTM are 170, 188, and 200, respectively. The numbers of parameters for the three ResLSTMs are 1.02, 1.51, and 2.03 million, respectively.

ResNet-BC: The ResNet_BC models are formed with basic residual blocks. As shown in Fig. 3(a), each residual block contains two 1-D causal dilated convolution units with an output size of \( d_f = 64 \), where each convolution unit is pre-activated by a layer normalisation followed by the ReLU activation function. The kernel size of each convolution units

---

Fig. 3: Proposed multi-branch ResNet (MB-ResNet) architecture. It is composed of a fully-connected first layer, FC, followed by \( N \) multi-branch residual blocks, and then a fully-connected output layer, O that employs sigmoid units. \( \odot \) represent the concatenation operation.
is $K = 3$. The dilation rate $d$ in each block is cycled from 1 to 16 (increasing by power of 2). By cascading 40, 60, and 80 basic residual blocks, we build three ResNet_BC models with 1.03, 1.53, and 2.03 million parameters, respectively, as baselines. This network architecture is also known as temporal convolutional network (TCN).

ResNet-BK: The ResNet_BK models are built with bottleneck residual blocks. As shown in Fig. 3(b), each residual block contains three 1-D causal dilated convolution units, where each convolution unit is pre-activated by a layer normalisation followed by the ReLU activation function. The first and third convolution units in a bottleneck residual block have a kernel size of 1, whilst the second unit has a kernel size of $k = 3$. The dilation rate $d$ is cycled from 1 to 16 (increasing by power of 2). By cascading 20, 30, and 40 bottleneck residual blocks, we build three ResNet_BK models size of 1.05, 1.51, and 1.98 million parameters, respectively.

DenseNet: Fig. 4 shows that each dense block consist of four dilated causal convolution units with a kernel size of $k = 3$, where each convolution unit is pre-activated by a layer normalisation followed by the ReLU activation function. The output size of each convolution unit is $d_f = 24$. The dilation rate $d$ is cycled from 1 to 16 (increasing by power of 2). As baselines, DenseNets of sizes 0.97, 1.48, and 2.10 million parameters are formed by cascading 7, 9, and 11 dense blocks, respectively.

In addition to the aforementioned models which are evaluated in Deep Xi framework, the proposed enhancement system is also compared with two widely known deep learning speech enhancement frameworks, LSTM-IRM proposed in [19] and bidirectional LSTM (BLSTM)-IRM. For these two frameworks, ideal ratio mask (IRM) is used as training target and 23 consecutive input frames (11 past frames, 1 current frames, and 11 future frames) are concatenated into a feature vector that used as the input of the network to estimate current frame of the mask. The network used in LSTM-IRM is composed of four LSTM layers, each with 1024 units, and a fully connected layers with 257 sigmoid units. The network used in BLSTM-IRM consists of four bidirectional LSTM layers, each with 512 units, and a fully connected layers with 257 sigmoid units.

C. Database

Training Set: Here, we give a detailed description of the clean speech and noise recordings used for training models in this experiment. For the clean speech recordings, we use the train-clean-100 set from the Librispeech corpus as training set [38], which includes 28,539 utterances spoken by 251 speakers. To conduct cross-validation experiments, 1000 clean speech recordings are randomly selected from the training set to construct the validation set. The employed noise recordings in the training set are taken from the following datasets: the QUT-NOISE dataset [39], the Nonspeech dataset [40], the Environmental Background Noise dataset [41], [42], and the noise set from the MUSAN corpus [43]. This gives a total of 1,295 noise recordings. All clean speech and noise recordings are single-channel, with a sampling frequency of 16 kHz (recordings with a sampling frequency higher than 16 kHz are downsampled to 16 kHz). A description of how the noisy speech signals are generated from the clean speech and noise recordings is given in the next subsection.

Test Set: For testing, recordings of four different noise sources are employed to form the test set. Two of the four noise recordings are of the real-world non-stationary noise sources, which includes street music noise (recording no. 26,270) from the Urban Sound dataset [44], and voice babble noise from the RSG-10 noise dataset [45]. Another two of the four noise recordings of real-world coloured noise sources, including factory and F16 noises from the RSG-10 noise dataset [45]. We randomly choose 10 clean speech recordings (without replacement) from the TSP speech corpus [46] for each of the four noise recordings. To construct the noisy speech signals, we mix the clean speech signals with a random section of the noise recording at five different SNR levels: ranging from -5 to 15 dB with a step of 5 dB. This constructed a test set of 200 noisy speech signals. All the noisy speech signals were single channel, with a sampling frequency of 16 kHz.

D. Training Details

The training details for all the models (MB-ResNet and baseline models) are described as follows:

- Cross-entropy as the loss function.
- We employ the Adam optimizer [47] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 0.001 for gradient descent optimisation.

![Fig. 4: The illustration of Densenet block. Each dense block includes four 1-D causal dilation convolution units with an output of size $d_f$. The kernel size and dilation rate are denoted by $k$ and $d$, respectively.](image-url)
TABLE I: Speech enhancement performance of different networks in terms of wideband PESQ metric. The highest PESQ score obtained at each condition and for each parameter size is highlighted with bold text.

| Network          | # params. x 10^6 | SNR level (dB) | Voice babble | Street music | F16 | Factory |
|------------------|------------------|----------------|--------------|--------------|-----|---------|
|                  |                  | -5 0 5 10 15   | -5 0 5 10 15 | -5 0 5 10 15 | -5 0 5 10 15 |
| Noisy speech     | –                | 1.04 1.07 1.14 1.35 1.71 | 1.04 1.06 1.11 1.27 1.58 | 1.03 1.06 1.11 1.25 1.52 | 1.04 1.04 1.09 1.24 1.54 |
| ResLSTM          | 2.02             | 1.08 1.19 1.43 1.91 2.44 | 1.08 1.18 1.37 1.70 2.14 | 1.12 1.27 1.54 1.87 2.28 | 1.06 1.22 1.49 1.87 2.32 |
| DenseNet         | 0.97             | 1.05 1.16 1.41 1.83 2.28 | 1.06 1.15 1.36 1.64 2.09 | 1.11 1.26 1.47 1.76 2.13 | 1.04 1.15 1.39 1.74 2.17 |
| ResNet-BC        | 1.03             | 1.07 1.18 1.42 1.87 2.34 | 1.09 1.20 1.43 1.75 2.21 | 1.13 1.31 1.57 1.89 2.29 | 1.07 1.23 1.50 1.89 2.35 |
| ResNet-BK        | 1.05             | 1.08 1.23 1.53 1.92 2.37 | 1.10 1.24 1.49 1.80 2.24 | 1.16 1.36 1.60 1.88 2.25 | 1.11 1.30 1.55 1.85 2.28 |
| Prop. MB-ResNet  | 1.05             | 1.09 1.25 1.55 2.04 2.57 | 1.11 1.26 1.52 1.87 2.38 | 1.16 1.37 1.65 2.03 2.46 | 1.12 1.29 1.57 1.94 2.42 |
| ResLSTM          | 1.51             | 1.07 1.19 1.46 1.90 2.44 | 1.10 1.20 1.39 1.70 2.18 | 1.08 1.26 1.51 1.87 2.24 | 1.06 1.20 1.46 1.80 2.29 |
| DenseNet         | 1.48             | 1.05 1.15 1.39 1.83 2.32 | 1.06 1.14 1.32 1.56 2.02 | 1.09 1.26 1.51 1.82 2.23 | 1.05 1.15 1.43 1.78 2.23 |
| ResNet-BC        | 1.53             | 1.06 1.20 1.46 1.85 2.31 | 1.09 1.21 1.44 1.75 2.18 | 1.14 1.27 1.52 1.84 2.22 | 1.10 1.27 1.53 1.88 2.29 |
| ResNet-BK        | 1.51             | 1.09 1.24 1.53 1.98 2.46 | 1.11 1.24 1.52 1.86 2.22 | 1.17 1.38 1.65 1.92 2.29 | 1.14 1.29 1.54 1.88 2.30 |
| Prop. MB-ResNet  | 1.43             | 1.10 1.25 1.57 2.01 2.50 | 1.13 1.28 1.53 1.81 2.30 | 1.21 1.42 1.67 2.04 2.44 | 1.15 1.33 1.58 1.96 2.41 |
| ResLSTM          | 2.03             | 1.08 1.20 1.48 1.96 2.50 | 1.09 1.20 1.42 1.76 2.24 | 1.09 1.25 1.50 1.79 2.17 | 1.08 1.23 1.50 1.87 2.37 |
| DenseNet         | 1.94             | 1.06 1.18 1.44 1.87 2.33 | 1.06 1.18 1.42 1.70 2.14 | 1.09 1.28 1.53 1.83 2.20 | 1.05 1.21 1.54 1.93 2.35 |
| ResNet-BC        | 2.03             | 1.06 1.19 1.42 1.86 2.31 | 1.07 1.19 1.41 1.72 2.20 | 1.11 1.27 1.50 1.82 2.22 | 1.07 1.23 1.47 1.80 2.20 |
| ResNet-BK        | 1.98             | 1.07 1.20 1.53 1.96 2.52 | 1.10 1.26 1.52 1.89 2.41 | 1.18 1.40 1.67 2.00 2.35 | 1.11 1.28 1.56 1.93 2.41 |
| Prop. MB-ResNet  | 1.66             | 1.11 1.27 1.58 2.05 2.54 | 1.14 1.28 1.52 1.91 2.33 | 1.21 1.43 1.72 2.13 2.51 | 1.18 1.34 1.61 1.97 2.42 |
| LSTM-IRM         | 53.9             | 1.06 1.13 1.31 1.65 2.12 | 1.06 1.14 1.30 1.55 2.04 | 1.08 1.19 1.36 1.60 1.97 | 1.04 1.10 1.25 1.55 2.00 |
| BLSTM-IRM        | 39.0             | 1.07 1.17 1.40 1.81 2.24 | 1.08 1.17 1.36 1.65 2.07 | 1.09 1.23 1.43 1.73 2.12 | 1.05 1.14 1.34 1.66 2.07 |
| Prop. MB-ResNet  | 1.66             | 1.11 1.27 1.58 2.05 2.54 | 1.14 1.28 1.52 1.91 2.33 | 1.21 1.43 1.72 2.13 2.51 | 1.18 1.34 1.61 1.97 2.42 |

E. Evaluation Metrics

In our experiments, two objective metrics are used to evaluate the objective quality and intelligibility of the enhanced speech. The wideband perceptual evaluation of speech quality (Wideband PESQ) metric [48] is used to obtain the mean opinion score of the objective speech listening quality. The PESQ score ranges from 0 to 5, with a higher PESQ score implying better speech quality. The short-time objective intelligibility (STOI) [49, 50] is used to evaluate the objective speech intelligibility. It is based on the correlation coefficient between the temporal envelopes of the clean speech and enhanced speech in short-time regions. The STOI score ranges from 0 to 1, with a higher STOI score indicating better speech intelligibility.

F. Results

The wideband PESQ scores of the enhanced speech signals produced by each of the networks (ResLSTM, DenseNet, ResNet-BC, ResNet-BK, and MB-ResNet) in Deep Xi framework are listed in Table I. The a priori SNR is estimated by each network and then employed in the square-root wiener filtering (SRWF) estimator to obtain the enhanced speech. The maximum PESQ score for each condition and each parameter size is highlighted in boldface. From Table I it can be observed that MB-ResNet performs best in terms of objective speech quality scores for most of the tested conditions. The performance superiority of MB-ResNet is demonstrated as MB-ResNet with a parameter size of 1.05 million provides 0.21 PESQ improvement over ResNet-BK with same parameter size for F16 at 15 dB. The MB-ResNet also shows a superiority in terms of model size as the MB-ResNet with a parameter size of 1.43 million attains 0.11 PESQ improvement over ResNet-BK with a parameter size of 1.51 million for Factory at 15 dB.

In Table II we present the STOI scores of the enhanced speech signals produced by each of the models. From Table II it is seen, similarly to Table I that MB-ResNet is able to attain the highest STOI score under most of the tested conditions. MB-ResNet demonstrates its superiority by the evaluation results: for voice babble at -5, 0, and 5dB, MB-ResNet with a parameter size of 1.66 million obtains 5%, 4.9%, and 2% STOI improvements over ResNet-BK with 1.98...
TABLE II: Speech enhancement performance comparisons of different networks in terms of STOI metric. The highest STOI score obtained at each condition and for each parameter size is highlighted with bold text.

| Network          | # params. x 10^6 | Voice Babble  | Street Music  | F16  | Factory |
|------------------|------------------|---------------|---------------|------|---------|
| Noisy speech     | -                | 60.2 72.4 83.0 90.7 95.5 | 59.0 70.9 81.9 90.3 95.6 | 60.4 71.8 82.4 90.5 95.7 | 57.8 69.9 80.9 89.2 94.5 |
| ResLSTM          | 1.02             | 61.0 75.8 86.8 93.4 96.7 | 63.0 77.1 86.7 92.8 96.3 | 66.1 78.4 87.1 92.9 96.6 | 62.1 77.4 87.0 92.7 96.2 |
| DenseNet         | 0.97             | 58.3 73.1 86.0 93.0 96.5 | 59.1 74.0 84.7 92.0 96.2 | 63.5 77.8 86.7 92.4 95.9 | 56.8 73.0 85.1 91.7 95.9 |
| ResNet-BK        | 1.03             | 60.8 75.5 87.2 93.4 96.6 | 64.0 77.8 87.1 93.4 96.8 | 67.2 79.4 87.9 93.4 96.9 | 61.1 76.9 87.1 92.8 96.3 |
| ResNet-BK        | 1.05             | 62.6 78.1 88.4 94.2 97.0 | 66.0 79.7 88.5 94.0 97.0 | 68.4 80.8 88.7 93.9 97.1 | 64.9 79.1 88.2 93.3 96.4 |
| Prop. MB-ResNet  | 1.05             | 62.1 78.1 88.6 94.3 97.1 | 66.8 80.0 89.2 94.3 97.2 | 69.2 82.2 89.7 94.5 97.4 | 64.7 80.0 88.3 93.4 96.5 |
| ResLSTM          | 1.51             | 59.6 74.6 86.3 93.0 96.6 | 62.9 76.9 86.8 92.8 96.4 | 64.0 77.4 87.4 93.1 96.3 | 61.0 76.6 86.7 92.6 96.2 |
| DenseNet         | 1.48             | 58.1 73.7 85.9 93.1 96.6 | 58.2 73.3 84.8 91.9 96.1 | 65.9 78.3 86.9 92.6 95.9 | 57.5 73.0 84.8 91.5 95.4 |
| ResNet-BK        | 1.53             | 59.8 76.5 87.4 93.5 96.7 | 65.0 77.9 88.0 93.4 96.7 | 64.9 77.6 87.0 93.4 96.9 | 63.3 79.1 87.6 92.8 96.3 |
| ResNet-BK        | 1.51             | 64.1 79.2 88.9 94.2 97.0 | 66.2 79.8 88.6 93.9 97.1 | 69.5 81.8 89.5 94.3 97.2 | 65.3 80.3 88.2 93.1 96.5 |
| Prop. MB-ResNet  | 1.43             | 64.7 79.4 89.1 94.4 97.2 | 69.2 81.7 89.5 94.3 97.2 | 71.0 83.1 90.1 94.8 97.5 | 67.4 80.7 88.4 93.3 96.6 |
| ResLSTM          | 2.03             | 62.7 76.6 87.3 93.7 96.9 | 65.9 78.7 87.9 93.6 96.9 | 65.8 79.3 87.6 93.3 96.6 | 62.4 77.2 87.0 92.7 96.4 |
| DenseNet         | 1.94             | 57.8 73.4 85.8 92.3 95.8 | 60.4 75.2 86.8 93.1 96.6 | 63.4 77.8 87.1 92.9 96.0 | 58.4 74.4 86.4 92.6 96.2 |
| ResNet-BK        | 2.03             | 60.3 76.3 87.4 93.6 96.7 | 63.1 78.0 87.4 93.3 96.8 | 67.6 78.9 87.5 93.5 97.1 | 60.5 77.6 87.2 92.7 96.1 |
| ResNet-BK        | 1.98             | 59.0 74.1 87.3 93.8 97.1 | 65.3 80.6 89.1 94.1 97.1 | 68.7 81.5 89.3 94.3 97.1 | 64.3 79.9 88.5 93.2 96.6 |
| Prop. MB-ResNet  | 1.66             | 64.0 79.0 89.3 94.5 97.2 | 69.5 81.7 89.3 94.2 97.1 | 70.6 82.7 90.0 94.7 97.4 | 67.0 81.2 88.6 93.3 96.6 |
| LSTM-IRM         | 53.9             | 63.0 76.0 85.6 92.2 95.9 | 64.7 76.5 85.5 91.7 95.9 | 69.4 79.6 87.1 92.6 95.9 | 61.2 75.5 85.1 91.6 95.6 |
| BLSTM-IRM        | 39.0             | 64.7 77.7 87.1 93.1 96.4 | 67.7 78.3 86.7 92.7 96.5 | 69.3 79.7 87.8 93.2 96.7 | 63.9 78.5 86.9 92.4 95.9 |
| Prop. MB-ResNet  | 1.66             | 64.0 79.0 89.3 94.5 97.2 | 69.5 81.7 89.3 94.2 97.1 | 70.6 82.7 90.0 94.7 97.4 | 67.0 81.2 88.6 93.3 96.6 |

G. Parameter Efficiency

For many real-world speech processing applications, memory resources and computational complexity are considerable constraints. For comparison of parameter efficiency, in Tables I and II we present the number of trainable parameters in different models. From the numbers in the tables, it can be seen that our proposed MB-ResNet exhibits higher parameter efficiency than ResLSTM, DenseNet, ResNet-BC, and ResNet-BK models. Compared to another two widely known enhancement methods, LSTM-IRM and BLSTM-IRM, the proposed model also demonstrates a significant superiority in terms of parameter efficiency for low-power and low-memory required applications. To be specific, the parameter sizes of LSTM-IRM (53.9 M) and BLSTM-IRM (39.0 M) are around 32.5 times and 23.5 times that of MB-ResNet (1.66 M), respectively.

V. COMPARISON WITH OTHER STATE-OF-THE-ART METHODS

In this section, to further demonstrate its superiority, we compare the proposed model with other state-of-the-art methods on a same publicly available dataset. A fair comparison is ensured since all the models are optimized by the authors on the exact same dataset. The brief descriptions for the baseline methods are provided in next subsection.

A. Baseline Methods

In this experiment, we conduct the performance evaluations of the proposed MB-ResNet model (in Deep Xi framework) in comparison with the following baseline methods from the literature:

- **Wiener** [5], a traditional statistic-based Wiener filtering method that is based on the *a priori* SNR estimator.
- **SEGAN** [51], a time-domain speech enhancement model, using generative adversarial networks (GAN) to directly reconstruct the clean waveform from noisy speech waveform.
- **Wavenet** [52], a non-causal Wavenet-based denoising model, operating on the raw waveform. It employs a
regression loss function ($L_1$ losses on both the speech waveform and the noise waveform prediction branches).

- **Wave-U-Net** [53], a one-dimensional adaptation of U-Net architecture for time-domain speech enhancement.

- **Deep Feature Loss** [54], also a time-domain denoising model that is trained with a deep feature loss from another acoustic environment classifier network.

- **MMSE-GAN** [55], a time-frequency (T-F) masking based method, using a modified GAN to predict the clean T-F representation. The objective function includes a GAN objective and a $L_2$ loss between the predicted and the clean T-F representation.

- **Metric-GAN** [56], a T-F masking based method, using a GAN to directly optimize generators based on one or multiple evaluation metric scores to speech enhancement.

- **MDPhD** [57], a hybrid speech enhancement method of time-domain and time-frequency domain.

### B. Database

To make a direct and fair performance comparison, we used the same publicly available dataset [58] used in several previous works. In this dataset, the clean speech recordings comprise of 30 speakers from the Voice Bank Corpus [59] – 28 speakers were chosen for training and the remaining 2 for testing. The noisy speech in training set is synthesized using a mixture of clean speech with 10 types of noise, two of which are artificially generated and 8 real noise recordings are from the Diverse Environments Multi-channel Acoustics Noise Database (DEMAND) [60]. With respect to the test set, 20 different noisy conditions are included: 5 distinct types of noise sources from the DEMAND database at one of 4 SNR levels each (2.5, 7.5, 12.5, and 17.5 dB). In total, this produced 842 test samples (approximately 20 different sentences in each condition per test speaker). Both speakers and noise conditions in test set are totally unseen during training process. As in previous methods, the original raw waveforms were downsampled from 48 kHz to 16 kHz for training and testing.

![Fig. 5: Magnitude spectrograms (log scale) of (a) clean speech and (b) noisy speech (clean speech was mixed with voice babble at -5 dB). Enhanced speech produced by (c) ResLSTM 2.03 M, (d) DenseNet 1.94 M, (e) ResNet-BC 2.03 M, (f) ResNet-BK 1.98 M, (g) LSTM-IRM 53.9 M, (h) BiLSTM-IRM 39.0 M, and (i) proposed MB-ResNet 1.66 M.](image)

**TABLE III: Comparison with other state-of-the-art speech enhancement models on the second publicly available data set.** Higher score (CSIG, CBAK, COVL, PESQ, and STOI) indicates better performance and the highest scores obtained for each evaluation measure are highlighted with bold text. For comparison of parameter efficiency, we present the number of trainable parameters in different models.

| Methods                     | # Parameters, (M) | Types      | CSIG  | CBAK  | COVL  | PESQ  | STOI  |
|-----------------------------|-------------------|------------|-------|-------|-------|-------|-------|
| Noisy Input                 | -                 | -          | 3.35  | 2.44  | 2.63  | 1.97  | 92    | (91.5) |
| Wiener [56], Scalart et al. 1996 | -                 | T-F domain | 3.23  | 2.68  | 2.67  | 2.22  | -     |
| SEGAN [51], Pascual et al. 2017 | 43.2 M (25.8 M)   | Time-domain| 3.48  | 2.94  | 2.80  | 2.16  | 93    |
| Wavenet [52], Rethage et al. 2018 | 6.34 M            | Time-domain| 3.62  | 3.23  | 2.98  | -     | -     |
| Wave-U-Net [53], Macartney et al. 2018 | 10.2 M        | Time-domain| 3.52  | 3.24  | 2.96  | 2.40  | -     |
| Deep Feature Loss [54], Germain et al. 2018 | 0.64 M         | Time-domain| 3.86  | 3.33  | 3.22  | -     | -     |
| MMSE-GAN [55], Soni et al. 2018 | 0.79 M (0.56 M)   | T-F domain | 3.80  | 3.12  | 3.14  | 2.53  | 93    |
| Metric-GAN [56], Fu et al. 2019 | 1.89 M (0.35 M)  | T-F domain | 3.99  | 3.18  | 3.42  | 2.86  | -     |
| MDPhD [57], Kim et al. 2018 | 6 M               | Hybrid     | 3.85  | 3.39  | 3.27  | 2.70  | -     |
| Proposed Multi-ResNet (DeepXi - SRWF) | -                 | T-F domain | 4.20  | 3.32  | 3.55  | 2.87  | 94    | (93.70) |
| Proposed Multi-ResNet (DeepXi - MMSE-STSA) | 1.66 M        | T-F domain | 4.21  | 3.36  | 3.57  | 2.91  | 94    | (93.70) |
| Proposed Multi-ResNet (DeepXi - MMSE-LSA) | -                 | T-F domain | 4.21  | 3.41  | 3.59  | 2.94  | 94    | (93.64) |
C. Results and Performance Metrics

In addition to PESQ and STOI metrics, another three composite measures are also exploited to evaluate the enhancement performance of the proposed models and state-of-the-art competitors. The three composite measures are:

- CSIG [61]: mean opinion score (MOS) predictor of signal distortion attending only to the speech signal (from 1 to 5).
- CBAK [61]: MOS predictor of the background-noise intrusiveness (from 1 to 5).
- COVL [61]: MOS predictor of the overall speech quality (from 1 to 5).

Table III presents the comparison results of these metrics on the second dataset. For all baseline methods, the best results that have been reported in literature are listed. The missing values in the table are because the results are not reported in the work. Here, the Deep Xi framework using proposed multi-branch ResNet (MB-ResNet) to estimate the a priori SNR is integrated into the SRWF, MMSE-STSA, and MMSE-LSA speech estimators.

It can be clearly observed in Table III that the MB-ResNet outperforms time-domain methods such as SEGAN [51], Wavenet [52], Wave-U-Net [53], and Deep Feature Loss (DFL) [54] in terms of all five measures by a comfortable margin. For example, MB-ResNet provides 0.25, 0.08, and 0.37 improvements over DFL for CSIG, CBAK, and COVL, respectively. Our method also shows large performance gain over T-F based methods like MMSE-GAN [55] and MetricGAN [56]. For example, the MB-ResNet provides 0.22, 0.23, 0.17, and 0.08 improvements over MetricGAN for CSIG, CBAK, COVL, and PESQ respectively. The proposed model also provides 0.41 PESQ improvements and about 1% STOI improvements over MMSE-GAN. In addition, our model provides great improvement over a hybrid method of time-domain and time-frequency domain, MDPhD [57]. These evaluation results significantly demonstrate the superiority of the proposed model.

The magnitude (log scale) spectrograms of enhanced speech produced by SEGAN, DFL, and MB-ResNet (Deep Xi-SRWF 1.66 M) are shown in Fig. 6(c)-(e), respectively. It can be observed that MB-ResNet (e) is able to achieve better trade-off between noise suppression and speech distortion. At the beginning segment of speech (around 0.035 s), SEGAN almost exhibits no noise suppression (Fig. 6(c)). In addition, multiple residual noise components can be seen in the spectrograms enhanced by both SEGAN and DFL (Fig. 6(d)).

D. Parameter Efficiency

As mentioned in Section IV-G, the parameter efficiency of models is a considerable constraints for many real-world speech applications. For this, in Table III we present the the number (in millions) of learnable parameters in different models. In Table III, the numbers of learnable parameters in different models listed are reported value in literature or computed from the code provided by the authors. Additionally, since GAN-based methods need to train both generator and discriminator networks, the listed values represent the number of parameters in generator and discriminator (in bracket), respectively. From the Table III we can find that our proposed model is able to provide higher parameter efficiency than most state-of-the-art (SOTA) methods (SEGAN, Wavenet, Wave-U-Net, Metric-GAN, and MDPhD). Although MMSE-GAN has less parameter than proposed model, the training instabilities of GAN models are still not completely understood. For DFL, an extra feature loss model needs to be pre-trained. One must note that our proposed model has a significant performance improvement accompanied by a modest increasing of number of trainable parameters compared to MMSE-GAN and DFL. In addition, note that we can adjust the parameter efficiency of MB-ResNet simply by altering the multi-branch dilated residual blocks. Compared to these SOTA models, the presented model is able to achieve better trade-off between performance and parameter efficiency.

VI. Conclusion and Discussion

In this study, we have presented a MB-ResNet model for single-channel speech enhancement. The proposed model utilizes the split-transform-aggregate design, and incorporates the 1-D causal dilated convolutions and identity residual connection. The split-transform-aggregate design demonstrates a strong representation power at low computational complexity.
The combination of dilated convolutions and residual learning builds large receptive fields, which enables our proposed model to capture very long effective history information to make a prediction. Specifically, large receptive fields enable model to learn the temporal dynamics of speech very well. The experimental results demonstrate that our proposed model outperforms many other advanced networks, such as ResLSTM, ResNet with basic structure (ResNet), ResNet with bottleneck structure (ResNet-BK), and DenseNet. Compared to two widely known deep learning methods, LSTM-IRM and BLSTM-IRM, the MB-ResNet in Deep Xi framework shows a significant superiority in terms of both performance and parameter efficiency.

Moreover, the comparison results with many state-of-the-art (SOTA) speech enhancement algorithms also demonstrates that our method is able to provide better enhancement performance in terms of widely used five objective metrics. For low-power and low-memory required applications of deep learning based speech enhancement methods, the number of parameters (parameter efficiency) in models is considerable constraint. It is crucial to achieve an optimal trade-off between parameter efficiency and speech enhancement performance of the model. However, most SOTA baseline models are not able to provide high parameter efficiency. The experimental results demonstrate that our proposed model is able to achieve a better trade-off than many SOTA speech enhancement methods.

REFERENCES

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, “DFT-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art,” Synthesis Lectures on Speech and Audio Processing, vol. 9, no. 1, pp. 1–80, Jan. 2013.
[2] P. C. Loizou, Speech Enhancement: Theory and Practice. Boca Raton FL: CRC Press, Feb. 2013.
[3] 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, 2018.
[4] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust., Speech, Signal Process., vol. 27, no. 2, pp. 113–120, 1979.
[5] S. Kamath and P. Loizou, “A multi-band spectral subtraction method for enhancing speech corrupted by colored noise,” in ICASSP, vol. 4, Citeseer, 2002, pp. 44 164–44 164.
[6] P. Scalart and J. V. Filho, “Speech enhancement based on a priori signal to noise estimation,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, May 1996, pp. 629–632.
[7] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” Proceedings of the IEEE, vol. 67, no. 12, pp. 1586–1604, 1979.
[8] Y. Ephraim and D. Malah, “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator,” IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[9] ——, “Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator,” IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.
[10] M. S. Kavalekar, I. K. Nielsen, J. B. Boldt, and M. G. Christensen, “Model-based Speech Enhancement for Intelligibility Improvement in Binaural Hearing Aids,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 27, no. 1, pp. 99–113, Jan. 2019.
[11] Q. Zhang, M. Wang, Y. Lu, L. Zhang, and M. Idrees, “A novel fast nonstationary noise tracking approach based on mmse spectral power estimator,” IEEE Trans. Audio, Speech, Lang. Process., vol. 88, pp. 41–52, 2019.
[12] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons, 2009.
[13] N. Mohammadiha, P. Smaragdis, and A. Leijon, “Supervised and unsupervised speech enhancement using nonnegative matrix factorization,” IEEE Trans. Audio, Speech, and Language Process., vol. 21, no. 10, pp. 2140–2151, 2013.
[14] D. Wang and G. J. Brown, Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley-IEEE Press, 2006.
[15] Y. Wang and D. Wang, “DNN-based classification-based speech separation,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 7, pp. 1381–1390, 2013.
[16] D. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in Speech separation by humans and machines. Springer, 2005, pp. 181–197.
[17] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 12, pp. 1849–1858, 2014.
[18] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” IEEE Signal processing letters, vol. 21, no. 1, pp. 65–68, 2013.
[19] J. Chen and D. Wang, “Long short-term memory for speaker generalization in supervised speech separation,” The Journal of the Acoustical Society of America, vol. 141, no. 6, pp. 4705–4714, 2017.
[20] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[21] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” IEEE/ACM Transactions on audio, speech, and language processing, vol. 22, no. 10, pp. 1533–1545, 2014.
[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[24] 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.
[25] A. Nicolson and K. K. Palwal, “Deep learning for minimum mean-square error approaches to speech enhancement,” Speech Communication, vol. 111, pp. 44 – 55, 2019.
[26] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” arXiv preprint arXiv:1511.07122, 2015.
[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.
[29] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[30] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.
[31] L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” arXiv preprint arXiv:1607.06450, 2016.
[32] Y. Luo and N. Mesgarani, “Tasnet: Surpassing ideal time-frequency masking for speech separation,” arXiv preprint arXiv:1809.07454, 2018.
[33] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in Proceedings of the 27th International Conference on International Conference on Machine Learning, ser. ICML’10. ACM, 2010, pp. 807–814.
[34] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. v. d. Oord, A. Graves, and K. Kavukcuoglu, “Neural machine translation in linear time,” arXiv preprint arXiv:1610.10099, 2016.
[35] J. Picone, “Signal modeling techniques in speech recognition,” Proceedings of the IEEE, vol. 81, no. 9, pp. 1215–1247, Sept 1993.
[36] X. Huang, A. Acero, and H.-W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.
[37] K. Paliwal and K. Wojcicki, “Effect of analysis window duration on speech intelligibility,” IEEE Signal Processing Letters, vol. 15, pp. 785–788, 2008.
[38] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2015, pp. 5206–5210.

[39] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, “The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms,” in Proceedings Interspeech 2010, 2010, pp. 3110–3113.

[40] G. Hu, “100 nonspeech environmental sounds,” The Ohio State University, Department of Computer Science and Engineering, 2004.

[41] D. Rethage, J. Pons, and X. Serra, “A wavenet for speech denoising,” in Proceedings of the 36th International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2016, pp. 2204–2208.

[42] F. Saki and N. Kehtarnavaz, “Automatic switching between noise classification and speech enhancement for hearing aid devices,” in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Aug 2016, pp. 736–739.

[43] D. Snyder, G. Chen, and D. Povey, “MUSAN: A music, speech, and noise corpus,” CoRR, vol. abs/1510.08484, 2015. [Online]. Available: http://arxiv.org/abs/1510.08484

[44] J. Salamon, C. Iacob, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in 22nd ACM International Conference on Multimedia (ACM-MM’14), Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[45] H. J. Steeneken and F. W. Geurtsen, “Description of the RSG-10 noise database,” Report IZF 1988-3, TNO Institute for Perception, Soesterberg, The Netherlands, 1988.

[46] P. Kabal, “TSP speech database,” McGill University, Database Version, 2002.

[47] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” CoRR, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[48] “Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs,” ITU-T Recommendation P. 862.2.

[49] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, March 2010, pp. 4214–4217.

[50] ——, “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125–2136, Sept 2011.

[51] S. Pascual, A. Bonafonte, and J. Serra, “SEGAN: speech enhancement based on generative adversarial network,” in Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH), Stockholm, Sweden, Aug. 2017, pp. 3642–3646.

[52] D. Rethage, J. Pons, and X. Serra, “A wavenet for speech denoising,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, April 2018, pp. 5069–5073.

[53] C. Macartney and T. Weyde, “Improved speech enhancement with the Wave-U-Net,” arXiv preprint arxiv: 1811.11307, 2018.

[54] F. G. Germain, Q. Chen, and V. Koltun, “Speech denoising with deep feature losses,” arXiv preprint arxiv: 1806.10522, 2018.

[55] M. H. Soni, N. Shah, and H. A. Patil, “Time-frequency masking-based speech enhancement using generative adversarial network,” in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, April 2018, pp. 5039–5043.

[56] S. Fu, C. Liao, Y. Tsao, and S. Lin, “MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement,” in Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, USA.

[57] J.-H. Kim, J. Yoo, S. Chun, A. Kim, and J.-W. Ha, “Multi-domain processing via hybrid denoising networks for speech enhancement,” arXiv preprint arxiv:1812.08914, 2018.

[58] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, “Investigating rnn-based speech enhancement methods for noise-robust text-to-speech,” in Proceedings of the 9th Speech Synthesis Workshop (ISCA), Sunnyvale, USA.

[59] C. Veaux, J. Yamagishi, and S. King, “The voice bank corpus: Design, collection and data analysis of a large regional accent speech database,” in 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE). IEEE, 2013, pp. 1–4.