INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Convolutional Neural Network with Non-Local Module for Speech Enhancement

Xiaoqi Li, Yaxing Li, Meng Li, Shan Xu, Yuanjie Dong, Xinrong Sun, Shengwu Xiong*

School of Computer Science and Technology, Wuhan University of Technology, Wuhan, Hubei, China

* whhit173@hotmail.com

Abstract

Convolution neural networks (CNNs) are achieving increasing attention for the speech enhancement task recently. However, the convolutional operations only process a local neighborhood (several nearest neighboring neurons) at a time across either space or time direction. The long-range dependencies can only be captured when the convolutional operations are applied recursively, but the problems of computationally inefficient and optimization difficulties are introduced. Inspired by the recent impressive performance of the non-local module in many computer vision tasks, we propose a convolutional neural network with non-local module for speech enhancement in this paper. The non-local operations are capable of capturing the global information in the frequency domain through passing information between distant time-frequency units. The non-local operations are able to set the dimension of the input as an arbitrary value, which results in the easy integration with our proposed network framework. Experimental results demonstrate that the proposed method not only improves the computational efficiency significantly but also outperforms the competing methods in terms of objective speech intelligibility and quality metrics.

Index Terms: speech enhancement, convolutional neural network, non-local module

1. Introduction

In most applications, speech enhancement is concerned with improving the quality and intelligibility of speech that has been degraded by additive noise [1]. Speech enhancement algorithms not only improve some perceptual aspect of speech but also can be used as a preprocessing step for other speech-related tasks, such as automatic speech recognition [2], speech coding [3] and hearing aids [4]. Traditional speech enhancement methods have been studied for decades, and most of these methods are based on the additional nature of background noise or the statistical properties of speech and noise signals including spectral subtraction [5], iterative wiener filtering [6], and minimum mean-square error of the spectra (MMSE) algorithms [7]. However, for highly non-stationary noise scenarios, these statistical-based methods usually fail to build estimators and therefore introduce additive artifacts in the enhanced speech due to the unrealistic assumptions [1].

In the past few years, deep learning based speech enhancement approaches have been extensively investigated and shown to provide a significant performance improvement over conventional statistical-based methods [8-12]. Xu *et al.* introduced a regression-based speech enhancement framework using deep neural networks (DNNs) with a multiple-layer deep architecture and DNNs are used as the regression model to predict the clean speech log power spectrum (LPS) from the noisy LPS features [11]. In [10], the masking-based approaches learn a mapping function from noisy speech features to a time-frequency mask and the estimated speech signal is obtained as the product of the noisy speech features and estimated time-frequency mask. The multitask learning approaches use a neural network to jointly estimate the primary target and other secondary features for speech enhancement. Gao *et al.* propose a joint framework combining speech enhancement and voice activity detection (VAD) to increase the speech intelligibility in harsh environments [8]. Except the learning targets, the supervised speech enhancement methods also are investigated from the aspects of input feature and network structure. Some researchers operate at the waveform level, training the model end-to-end, other than spectral domain or some higher-level features. The time domain waveform enhancement frameworks based on generative adversarial networks (GANs) [13, 14], fully convolutional neural network [15-17] and WaveNet [18, 19] have been introduced. The recurrent neural network (RNN) [20, 21] was investigated to capture the temporal dependences of speech signal and significantly outperforms the DNN with feedforward structure. Compared to RNN, CNN improves the computation efficiency due to its weight sharing property and some CNN-based structures have already proved to handle sequence problems well [22, 23]. Furthermore, the models based on convolutional and recurrent neural networks (C-RNN) for speech enhancement [24, 25] achieve better generalization on both seen and unseen noise.

Convolutional operations process a local neighborhood (several nearest neighboring neurons) across either space or time direction. The long-range dependencies can be captured when the convolutional operations are applied recursively, by which a larger receptive field is obtained from the previous layers of the network. If the long-range dependencies are captured in advance by some convolutional layers, both the performance and the computational efficiency of the network can be improved. To achieve this target, the dilated convolution [26] is exploited to enlarge the receptive field and increase the contextual information captured in the network layer. The fully-connected layer can get the global information

This work was supported by the National Key Research and Development Program of China (Grant No. 2017YFB1402203), the Defense Industrial Technology Development Program (Grant No. JCKY2018110C165), Hubei Provincial Natural Science Foundation of China (Grant No. 2017CFA012). This work was also supported by "the Fundamental Research Funds for the Central Universities". The Corresponding author is Yaxing Li.

http://dx.doi.org/10.21437/Interspeech.2019-2472

in one layer, but it brings a lot of parameters and increases the difficulty of network optimization. The fully-connected layer also requires a fixed-size input/output and loses positional correspondence. These shortcomings limit the use of the fully-connected layer in many cases. To address the above issues of convolutional operations, a non-local module [27] has been proposed recently and achieved impressive performance after the integration with CNNs in many computer vision tasks [27-30]. In this work, we propose a convolutional neural network with the non-local module for speech enhancement. The non-local operations calculate the mutual similarity between time-frequency units in each frame, which is helpful for capturing the global information in the frequency domain with a slightly increased computing complexity. The experimental results show that the proposed scheme improves the computational efficiency significantly and produces satisfactory enhancement performance comparing the DNN, long short-term memory (LSTM) and C-RNN baselines.

The rest of this paper is organized as follows. In section 2, we describe the non-local module and proposed network structure in detail. Section 3 presents the experimental methods, experimental results and analysis. The summary and conclusions are given in section 4.

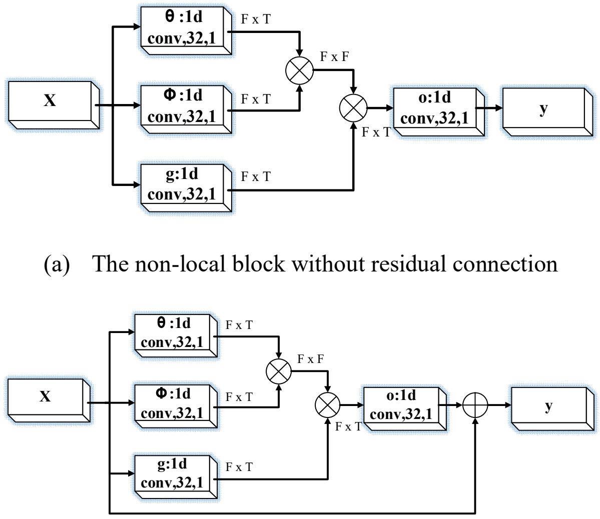
2. Non-local convolutional neural network for speech enhancement

2.1. Non-Local block for speech enhancement

The non-local operation aggregates the information from the inputs based on their similarity and is defined as follows:

$$\mathrm{y}_i = \frac{1}{\mathcal{C}(x)} \sum_{\forall j} f(x_i, x_j) g(x_j), \tag{1}$$

where x and y, respectively, denote the input and output tensor of the operation with the same size, f represents the pairwise function to calculate the correlation between the locations of the feature map, g signifies the unary input function for information transform, and $\mathcal{C}(x)$ is a normalization factor. The similarity functions such as the embedded Gaussian, dot-product and concatenation have been introduced [27]. The self-attention module recently presented for machine translation is found to be a special case of non-local operations in the embedded Gaussian version. Thus, we select the embedded Gaussian similarity function given as follows:



(a) The non-local block without residual connection

(b) The non-local block with residual connection

Figure 1: *The structure of non-local blocks*

$$f(x_i, x_j) = e^{\theta(x_i)^T \phi(x_j)}, \tag{2}$$

where $\theta(x_i)$ and $\phi(x_j)$ are two embedding schemes using a convolution with kernel size 1. In this case the normalization factor in Eq. (1) is set as $\mathcal{C}(x) = \sum_{\forall j} f(x_i, x_j)$ and $\frac{f(x_i,x_j)}{\mathcal{C}(x)}$ becomes the softmax operation along the dimension j.

The input and output of the non-local operation have the same number of arbitrary determined dimensions. Therefore, the non-local operation can be easily combined with other architectures to construct a network block. In this paper, we consider using the 1D convolution, in which the time direction is regarded as the channel dimension and the convolution kernel performs the convolution operation across the frequency dimension. The non-local blocks are capable of passing information between distant time-frequency units based on their similarity, and thus can be applied for the speech enhancement task. The non-local blocks without and with residual connection are defined as follows:

$$z_i = o(y_i), \tag{3}$$

$$z_i = o(y_i) + x_i, \tag{4}$$

where $o(\cdot)$ represents the convolution operation without a bias item, y_i is calculated using the non-local operation in Eq. (1), and the addition of x_i denotes a residual connection [31]. The introduced non-local blocks without and with residual connection are illustrated in Figure 1.

The dimension of the introduced non-local blocks are specified as $F{\times}T$, where F and T denote the numbers of frequency channels and time frames in the short-time Fourier transform (STFT) magnitude spectra, respectively. The function $f \in \mathrm{R}^{F\times F}$ calculates the similarity between each time-frequency unit in the same frame based on $\theta(\cdot) \in \mathrm{R}^{F\times T}$ and $\phi(\cdot) \in \mathrm{R}^{F\times T}$. The softmax operation $\frac{f(x_i,x_j)}{\mathcal{C}(x)}$ in Eq. (1) is then processed along the frequency dimension. The non-local operation output y is calculated by the matrix multiplication of f and $g \in \mathrm{R}^{F\times T_c}$, where T_c is the number of channels obtained after dimension reduction in time domain. Finally, the output of the non-local block z is the spectrum information recovered by Eq. (3) or Eq. (4). Since the input x and output z have the same arbitrarily determined dimensions, the non-local block can be intergraded with existing mainstream networks easily.

2.2. Network configurations

Figure 2 illustrates the proposed network architecture with non-local blocks for speech enhancement. The extension block expands the number of time and frequency channels, which enable the use of larger receptive field in the following layers. Subsequently, we use multiple layers of convolutional layers and some of which are connected with the non-local block. The 1-D convolutional layer is then employed for dimension reduction and the final estimated clean speech feature is obtained by a fully-connected layer with linear activation. The

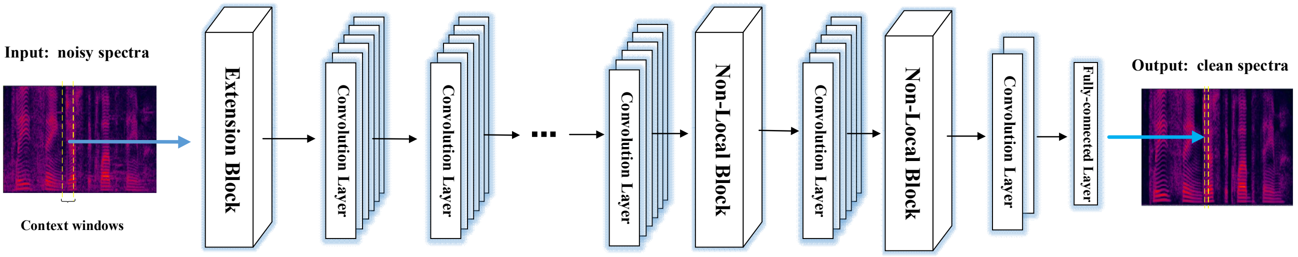


Figure 2: *Overview of the proposed architecture for speech enhancement*

Table 1: *Network configurations of the proposed architecture. The layer parameters are given in (outChannels, kernelSize, Stride) format.*

Layers		Input size	Parameters		Output size
Input-layer		$(T,129)$	~		~
Extension Block	FD conv	$(T,129)$	(32,1,1)		(256,32)
	TD conv		(256,3,1)		
Convolution Layer		(256,32)	(32,3,1)	X 4	(256,32)
Convolution Layer		(256,32)	(32,3,1)	X 2	(256,32)
Non-Local Block	θ: conv		(32,1,1)		
	ϕ: conv		(32,1,1)		
	g: conv		(32,1,1)		
	o: conv		(32,1,1)		
Convolution Layer		(256,32)	(2,1,1)		(256,2)
Flatten		(256,2)	~		(512,)
Fully-connected Layer		(512,)	(129,)		(129,)

activation functions of all the convolutional layers in our proposed network are ELUs [32]. Table 1 shows the detailed description of our proposed network architecture.

The first few layers of the network (several convolutional layers before non-local block) are used to process the locally time-frequency information and the well learned high level features can be passed to the subsequent network layers. The computational load of the non-local block is lightweight if it is used in the high-level feature maps. We also find that the simply increase of the non-local blocks doesn't achieve better speech enhancement performance. Therefore, the two non-local blocks are inserted into the last three layers of convolutional layers, as shown in Figure 2.

3. Experiments

The experiments are conducted on TIMIT database [33]. A total of 1000 sentences are randomly selected for training and another 400 sentences excluded from the training speech are used to construct the testing set. Babble, factory1 and white Gaussian noise from the NOISEX-92 database [34] and railway noise from the Aurora2 database [35] are used as noise signals. A noise segment with same length as clean speech utterance from the abovementioned four noise types is randomly picked and mixed with clean speech to generate a set of artificially noisy utterances with signal-to-noise-ratios (SNRs) from -5 to 15 dB, with 5 dB increments. For the signal analysis, the original raw waveforms are firstly down-sampled to 8 kHz and a 256-point Hamming window is applied with a 50% overlap. The noisy and clean speech spectra are represented by the 129 dimensional LPS features. The input and output features are standardized to zero mean and unit variance, and a reverse step is processed on the output. We randomly select 10% of the LPS features as the validation set to prevent overfitting during network training. In order to measure the robustness of different speech enhancement models, we additionally added the test with mismatched noise and mismatched SNR. The factory2 noise from the NOISEX-92 database and two other noise types from the Aurora2 database, namely restaurant and street, are used for mismatch evaluation. Two unseen SNR levels with -3 dB and 3 dB are also used for performance evaluation under noise match and noise mismatch conditions. For all experiments, we used the same experimental setups in order to perform direct performance comparison. We compare our proposed method with the following three baselines:

1. DNN [11]. DNN contains 3 hidden layers of size 1024 and sigmoid activation functions are applied to all fully-connected layer except the output layer.

2. LSTM [20]. LSTM contains 2 hidden layers, both of which has 1024 hidden units. And the outputs of LSTMs are feed into one fully-connected layer.

3. C-RNN [25].The convolutional component uses a 2-D convolution with 64 filters, kernel size (T, 16) and time-frequency stride (1, 8). Two layers of bidirectional LSTMs follow the convolutional component, each of which has 512 hidden unit.

Table 2: *PESQ and STOI scores of different numbers of non-local blocks in noise match condition.*

Metrics	PESQ					STOI				
Method	Noisy	+1 NL	+2 NL	+3 NL	+4 NL	Noisy	+1 NL	+2 NL	+3 NL	+4 NL
babble	2.0919	2.4979	**2.5073**	2.4757	2.4263	0.7166	0.7664	**0.7812**	0.7718	0.7680
factory1	1.9882	2.5712	**2.5829**	2.5646	2.5542	0.7101	0.7788	**0.7929**	0.7880	0.7887
railway	2.0519	**2.6237**	2.5823	2.5763	2.5512	0.7254	0.7993	**0.8065**	0.8043	0.8040
white	1.7270	2.5753	**2.6120**	2.5334	2.5438	0.7111	0.7752	**0.7869**	0.7829	0.7839
Avg.	2.1982	2.5670	**2.5711**	2.5375	2.5189	0.7120	0.7799	**0.7919**	0.7868	0.7861

Table 3: *PESQ and STOI values of the different methods on all noises types.*

Metrics	PESQ						STOI					
Method	Noisy	DNN	LSTM	C-RNN	NL	Res-NL	Noisy	DNN	LSTM	C-RNN	NL	Res-NL
SNR 15	2.7007	2.8067	3.0581	3.0749	3.2441	**3.2799**	0.9252	0.8675	0.9007	0.8932	**0.9224**	0.9198
SNR 10	2.3713	2.6743	2.8930	2.9245	3.0209	**3.0486**	0.8571	0.8432	0.8784	0.8724	**0.8954**	0.8930
SNR 5	2.0415	2.4715	2.6540	2.6979	2.7556	**2.7662**	0.7611	0.7967	0.8355	0.8334	**0.8485**	0.8452
SNR 3	1.9110	2.3601	2.5310	2.5816	2.6256	**2.6344**	0.7171	0.7696	0.8095	0.8110	**0.8225**	0.8187
SNR 0	1.7254	2.1744	2.3173	2.3900	2.4180	**2.4228**	0.6463	0.7164	0.7561	0.7663	**0.7720**	0.7672
SNR -3	1.5503	1.9496	2.0619	2.1631	2.1720	**2.1797**	0.5749	0.6490	0.6829	0.7067	**0.7082**	0.7027
SNR -5	1.4532	1.7904	1.8808	1.9923	1.9937	**2.0148**	0.5290	0.5976	0.6239	**0.6581**	0.6557	0.6527
Avg.	1.9648	2.3181	2.4852	2.5463	2.6043	**2.6209**	0.7158	0.7486	0.7839	0.7916	**0.8035**	0.7999

The input context window of each method is set as 11 ($T = 11$) and it spans from past 5 frames to future 5 frames. We use Adam optimizer [36] for the training and the initial learning rate is set to 0.001 with $\beta_1 = 0.9$，$\beta_2 = 0.999$ and $\epsilon = 1.0e^{-8}$. The mini-batch size is set to $N = 128$ and the loss function is mean square error (MSE). Total number of epoch is 100. When the validation loss doesn't decrease for more than 5 epochs, the training will be terminated in advance to prevent overfitting. We select the model which produces the best performance on the validation set for experiments. The objective speech quality and intelligibility is evaluated via perceptual evaluation of speech quality (PESQ) [37] and Short-Time Objective Intelligibility (STOI) [38] scores, respectively.

We firstly determine the optimal non-local blocks number. The average PESQ and STOI values of the enhanced speech with different numbers of non-local blocks on the test set at seven SNR levels (-5, -3, 0, 3, 5, 10, 15 dB) across four match noise types are given in Table 2. We find that the best performance has been achieved when two non-local blocks are exploited. It can be found that more than two non-local blocks fed in network does not results in better performance. It also indicates that more attention should be payed to the local time-frequency information of the first few layers to learn better features for the following layers of the network.

Table 3 illustrates the average PESQ and STOI values of the baseline methods and the proposed method with two non-local blocks across seven SNR levels for all noises types. We can clearly find that the proposed method consistently outperforms better than DNN and LSTM approaches. Our method produces comparable or better performance compared with C-RNN, and notable performance advantage is evident under high SNR levels. We also notice that the non-local block without residual connection produces a little worse PESQ values and slightly higher STOI in comparison with the structure with residual connection. The enhanced spectrograms from one noisy speech utterance corrupted by babble noise at SNR=3 dB using different methods are shown in Figure 3. It is observed that the spectrogram enhanced by our approach preserves the structure of the original signal better than the C-RNN, especially in the portion marked with the yellow rectangular boxes. Table 4 shows the number of parameters between the baselines and proposed model and it reveals that our proposed method achieves a large boost on the computational efficiency.

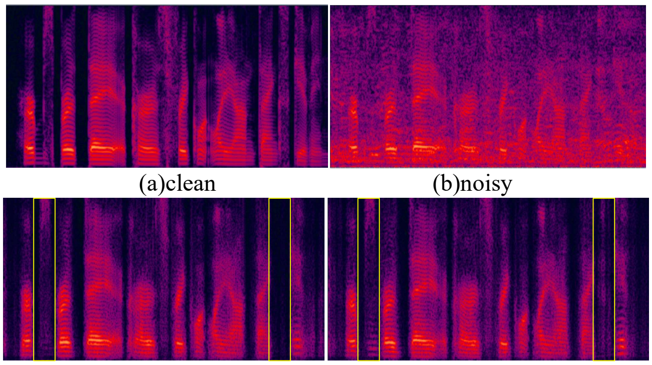

(a)clean (b)noisy
(c)C-RNN (d) NL

Figure 3: *Four spectrograms of an utterance corrupted by babble noise at SNR=3dB: (a) clean speech (b) noisy speech (c) enhanced by C-RNN (d) enhanced by our network with non-local block.*

Table 4: *The number of parameters between the baselines and proposed model.*

Method	Number of parameters (Million)
DNN	3.68
LSTM	13.25
C-RNN	12.47
NL	0.13

4. Conclusions

A new convolutional neural network with non-local block is proposed for speech enhancement in this paper. The non-local operations are capable of capturing the long-range dependencies in the frequency domain by calculating the mutual similarity between time-frequency units. The non-local block is a flexible building block which can be easily used together with any existing architectures. Our experimental results shows that the introduced architecture produces consistently better enhancement performance than other baselines. Additional experimental results indicate that more attention should be payed to the local time-frequency information of the first few layers to learn better features for the following layers of the network. In the future, we will try to apply a more efficient approach to aggregates the global information of the 2D spectrograms in the non-local module for speech enhancement.

5. References

[1] P. C. Loizou, Speech Enhancement: Theory and Practice: CRC Press, Inc., 2007.

[2] T. Ochiai, S.Watanabe, T. Hori, and J. Hershey, “Multichannel end-to-end speech recognition,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 2632-2641.

[3] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, “Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication,” *Speech Communication*, vol. 53, no. 5, pp. 677–689, 2011.

[4] A. Chern, Y. Lai, Y. Chang, Y. Tsao, R. Y. Chang, and H. Chang, “A smartphone-based multi-functional hearing assistive system to facilitate speech recognition in the classroom,” *IEEE Access*, vol. 5, pp. 10339–10351, 2017.

[5] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *Acoustics Speech & Signal Processing IEEE Transactions* on, vol. 27, no. 2, pp. 113-120, 1979.

[6] 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, 2005.

[7] Y. Ephraim, and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans Acoust Speech Signal Process*, vol. 32, no. 6, pp. 1109-1121,2003.

[8] T. Gao, J. Du, Y. Xu, C. Liu, L. R. Dai, and C. H. Lee, “Improving Deep Neural Network Based Speech Enhancement in Low SNR Environments,” *in International Conference on Latent Variable Analysis and Signal Separation,* 2015, pp. 75-82.

[9] Y. Zhao, D. L. Wang, I. Merks, and T. Zhang, “DNN-based enhancement of noisy and reverberant speech,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016,*2016, pp.6525–6529.

[10] Y. Wang, A. Narayanan, and D. L. Wang, “On Training Targets for Supervised Speech Separation,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 22, no. 12, pp. 1849-1858, 2014.

[11] 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, 2014.

[12] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, “A Regression Approach to Speech Enhancement Based on Deep Neural Networks,” *IEEE/ACM Transactions on Audio Speech & Language Processing*, vol. 23, no. 1, pp. 7-19, 2015.

[13] S. Pascual, A. Bonafonte, and J. Serrà, “SEGAN: Speech Enhancement Generative Adversarial Network,” in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association,* 2017, pp.3642–3646.

[14] D. Michelsanti and Z. Tan, “Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification,” in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association,* 2017, pp. 2008–2012.

[15] K. Tan, J. Chen, and D. L. Wang, “Gated Residual Networks with Dilated Convolutions for Supervised Speech Separation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018,* 2018, pp. 21–25.

[16] S. R. Park, and J. Lee, “A Fully Convolutional Neural Network for Speech Enhancement,” in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association,* 2017, pp. 1993–1997.

[17] S. W. Fu, Y. Tsao, X. Lu, and H. Kawai, “Raw Waveform-based Speech Enhancement by Fully Convolutional Networks,” in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference,* 2017, pp. 6–12.

[18] D. Rethage, J. Pons, and X. Serra, “A Wavenet for Speech Denoising,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018,* 2018, pp. 5069–5073

[19] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, “Speech Enhancement Using Bayesian Wavenet,” in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association,* 2017, pp. 2013–2017.

[20] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. W. Schuller, “Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR,” in *Latent Variable Analysis and Signal Separation - 12th International Conference,* 2015, pp. 91–99.

[21] J. Chen, and D. Wang, “Long short-term memory for speaker generalization in supervised speech separation,” in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association,* 2016, pp. 3314–3318.

[22] Y.N. Dauphin, A.Fan, M.Auli, D.Grangier, “Language Modeling with Gated Convolutional Network”, in *Proceedings of the 34th International Conference on Machine Learning, ICML 2017,* 2017,*pp. 933– 941.*

[23] P. Chen, W. Guo, Z. Chen, J. Sun, L. You, “Gated Convolutional Neural Network for Sentence Matching”, in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association,* 2018, pp. 2853–2857

[24] K. Tan and D. Wang, “A convolutional recurrent neural network for real-time speech enhancement,” in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association,* 2018, pp.3229–3233.

[25] H. Zhao, S. Zarar, I. Tashev, and C. Lee, “Convolutional-recurrent neural networks for speech enhancement,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018,* 2018, pp. 2401–2405.

[26] F. Yu and V. Kolt un, “Multi-scale context aggregation by dilated convolutions,” in *ICLR*, 2016.

[27] X. Wang, R. B. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *2018 IEEE Conference on Computer Vision and Pattern Recognition,* 2018, pp. 7794–7803.

[28] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, “Non-local recurrent network for image restoration,” in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018,* 2018, pp. 1680–1689.

[29] H. Levi, S. Ullman,”Efficient Coarse-to-Fine Non-Local Module for the Detection of Small Objects”, *arXiv preprint arXiv: 1811.12152,*2018.

[30] D. Y. Park and K. H. Lee, “Arbitrary Style Transfer with Style-Attentional Networks”, *arXiv preprint arXiv: 1812.02342,* 2018.

[31] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition,* 2016, pp. 770–778.

[32] D. Clevert, T. Unterthiner, S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus)”, in *ICLR*,2016

[33] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1,” *NASA STI/Recon Technical Report N*, vol. 93, Feb. 1993.

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

[35] H. G. H. a. D. Pearce, “The AURORA experimental framework for the preformance evaluations of speech recognition systems under noisy conditions,” in *ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium*, pp. 181–188.

[36] D. P. Kingma, and J. Ba, “Adam: A Method for Stochastic Optimization,” in *ICLR*, 2015.

[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 *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001,* 2001, pp. 749-752 vol.2.

[38] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “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, 2011.