DENSELY CONNECTED NETWORK WITH TIME-FREQUENCY DILATED CONVOLUTION FOR SPEECH ENHANCEMENT

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

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

ABSTRACT

The data driven speech enhancement approaches using regression-based deep neural network usually result in enormous number of model parameters, which increase the computational load and the difficulty of model training. In order to improve the model efficiency, we propose a densely connected network with time-frequency (*T-F*) dilated convolution for speech enhancement. The *T-F* dilated convolution block is designed to enlarge the receptive field and capture the contextual information in both temporal and frequency domains. Considering the computational efficiency, the 1-D convolution with the bottleneck structure is exploited in the *T-F* convolution block. Each *T-F* convolution block is then densely connected to ensure maximum information flow between layers and alleviate the vanishing gradient problem of the network. The experimental results reveal that the proposed scheme not only improves the computational efficiency significantly but also produces satisfactory enhancement performance comparing the competing methods.

Index Terms— Dense connectivity, dilated convolution, speech enhancement

1. INTRODUCTION

Speech enhancement has attracted considerable attention for several decades in the speech signal processing community due to its importance in applications such as speech communication, digital hearing aids and robust automatic speech recognition systems [1]. Based on the assumption on particular probabilistic models of speech and noise, statistical based methods including spectral subtraction [2], Wiener filtering [3], and minimum mean-square error of the spectra (MMSE) algorithms [4] have been proposed. However, for highly non-stationary noise scenarios, these statistical-based methods usually fail to build estimators and therefore introduce additional artifacts in the enhanced speech due to the unrealistic assumptions [1, 5].

In the past few years, the data driven or supervised approaches using regression-based deep neural network (DNN) have been shown to provide a significant performance improvement over conventional statistical-based methods [5-10]. Based on a definition of the learning targets, the data driven approaches are categorized as spectral mapping [6, 7] time-frequency (*T–F*) masking [11-13] and multitask learning methods [10, 14]. Xu *et al.* introduced a mapping-based approach and DNN is used as the regression model to predict the clean speech log power spectrum (LPS) from the noisy LPS features [6]. In [12, 13], the masking-based approaches learn a mapping function from noisy speech features to a *T–F* mask and the estimated speech signal is obtained as the product of the noisy speech features and estimated *T–F* mask. The multitask learning approaches use a neural network to jointly estimate the primary target and other secondary features for speech enhancement, by which additional constraints not available in the direct prediction are imposed and the learning of the primary target can be potentially improved [10, 14]. Except the learning targets, the supervised speech enhancement methods also are investigated from the aspects of input feature and network structure. The time domain waveform enhancement frameworks based on generative adversarial networks (GANs) [15], fully convolutional neural network [16-18] and WaveNet [19, 20] have been introduced. The long short-term memory (LSTM) network [13, 21] was investigated to capture the temporal dependences of speech signal and significantly outperforms the DNN with feed-forward structure. More recently, the more complex convolutional recurrent neural network (C-RNN) has been introduced for speech enhancement [5, 8].

Although the data driven speech enhancement methods using regression-based DNN show clear performance advantage, the high model complexity and vanishing gradient problem are introduced. In this work, we propose a densely connected network with time-frequency dilated convolution for speech enhancement. The dense connectivity concatenates the convolution output of all preceding layers as inputs, which ensures maximum information flow between layers in the network and alleviates the vanishing

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), and the Key Technical Innovation Project of Hubei Provence of China (Grant No. 2017AAA122). The Corresponding author is Shengwu Xiong* (xiongsw@whut.edu.cn).

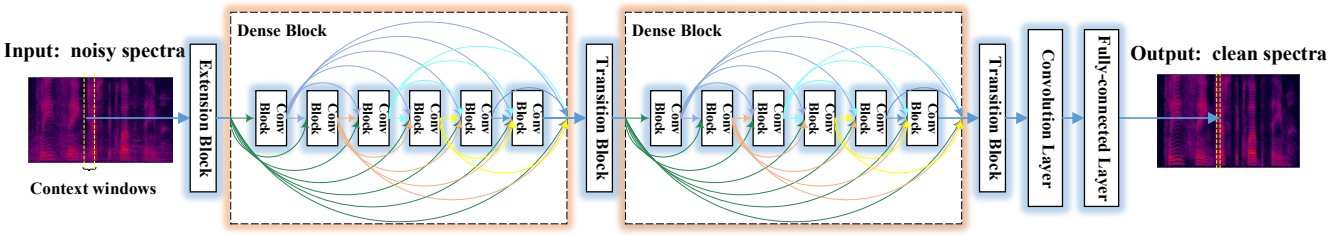

Fig. 1: Overview of the proposed architecture for speech enhancement

gradient problem. The time-frequency dilated convolutions with enlarged receptive field could capture the contextual information in both temporal and frequency domains. The introduced architecture is designed based on the fully convolutional framework and there is no recurrent layer used for temporal series modelling. The experimental results show that the proposed scheme improves the computational efficiency significantly and produces satisfactory enhancement performance comparing the DNN, LSTM and C-RNN baselines.

2. DENSELY CONNECTED NETWORK WITH TIME-FREQUENCY DILATED CONVOLUTION

A densely connected convolutional architecture, illustrated in Fig. 1, is explored to capture the contextual information in time and frequency for speech enhancement. A shape of $T\times F$ noisy spectra is fed into the network to estimate the corresponding clean speech spectra in the central frame, where T and F represents the input context frames and frequency channels, respectively. Due to the imbalance of T and F, the contextual information in frequency and time direction is aggregated separately. To improve the computational efficiency, the 1-D convolution is used instead the 2-D convolution.

2.1. Dense block

The input and the *l-th* layer output of the dense block is denoted as $\boldsymbol{x}_0$ and $\boldsymbol{x}_l$, respectively. The l-layer of the dense block receives the convolutional output of all preceding layers as the input

$$\boldsymbol{x}_l = H_l([\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{l-1}]), \quad (1)$$

where $H_l(\cdot)$ represents the non-linear transformation of the *l-th* layer, [...] denotes the concatenation of the convolutional output in the temporal direction, i.e. the concatenation of time channels. In [22], the transformation function $H_l(\cdot)$ is defined as the composite function of batch normalization (BN), followed by a rectifier non-linearity (ReLU) activation and a convolution layer. In our architecture, we design a *T-F* dilated convolutional block as the transformation function to capture the contextual information in both temporal and frequency domains. As shown in Fig. 1, there are six *T-F* dilated convolution blocks with a specific dilation rate each in one dense block. The *l-th* layer (*T-F* dilated convolution block) has $k_0+k\times(l-1)$ input time channels and the function $H_l(\cdot)$ produces k time channels, where k_0 is the number of time channels of the input $\boldsymbol{x}_0$ and the hyper-parameter k is referred to as the growth rate of the network. The growth rate for each dense block is set as a relatively small integer (k=16) to prevent the block growing too wide. The output of each *T-F* dilated convolution block should have the same number of frequency channels to ensure the concatenation of time channels.

2.2. *T-F* dilated convolutional block

The conventional convolutional neural networks (CNNs) do not capture the long-term temporal dependencies of speech signal due to the limited respective fields. One way to expand the respective fields is to increase the network depth, which typically decreases computational efficiency and results in vanishing gradients. The dilated convolution could expand the receptive fields while maintaining the network depth and the kernel size [23]. Fig. 2 illustrates the

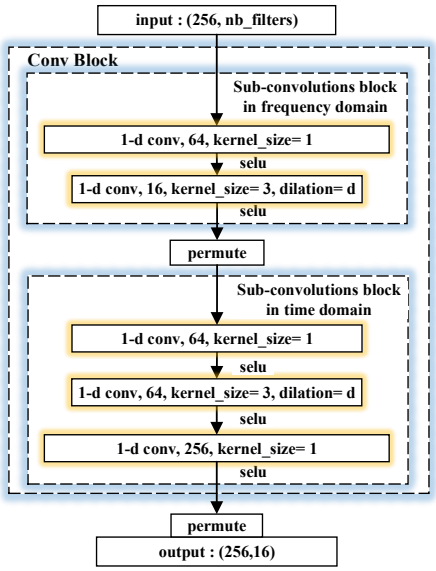

Fig. 2: The structure of *T-F* dilated convolutional block.

introduced dilated *T-F* convolutional block, in which the frequency and time sub-convolution blocks are connected sequentially. The 1-D frequency-dilated convolution with kernel size 3 is applied in the frequency sub-convolution block to capture the contextual information along the frequency direction. To improve the computational efficiency, the convolution with kernel size 1 can be introduced as an additional bottleneck layer before the dilated-convolution to reduce the number of input time channels. For the time sub-convolution block, the 1-D time-dilated convolution with kernel size 3 is exploited to capture the contextual information in temporal domain. An additional bottleneck layer with a fixed kernel size is introduced at the end the time-dilated convolution to make the output of each *T-F* convolutional block has a constant frequency channel. The self-normalizing neural networks with the scaled exponential linear unit (SELU) activation function allow the neuron activations converges to zero mean and unit variance automatically [24]. The network with SELUs activation results in notable better performance than that of the ReLU with BN. Hence, the SELU activation is exploited in the *T-F* convolution block instead of the ReLU with BN.

Table 1: Network configurations of the proposed architecture.

Layers			Input size	Parameter	Output size
Input-layer			(*T*,129)	~	~
Extension Block		FD conv	(*T*,129)	(32,1,1)	(32,256)
		TD conv		(256,3,1)	
Dense Block	Conv block	FD conv	(32,256)	(64,1,1) (16,3,1)	(128,256)
		TD conv		(64,1,1) (64,3,1) (256,1,1)	
	Conv block	FD conv		(64,1,1) (16,3,1)	
		TD conv		(64,1,1) (64,3,1) (256,1,1)	
	Conv block	FD conv		(64,1,1) (16,3,1)	
		TD conv		(64,1,1) (64,3,1) (256,1,1)	
	Conv block	FD conv		(64,1,1) (16,3,2)	
		TD conv		(64,1,1) (64,3,2) (256,1,1)	
	Conv block	FD conv		(64,1,1) (16,3,4)	
		TD conv		(64,1,1) (64,3,4) (256,1,1)	
	Conv block	FD conv		(64,1,1) (16,3,8)	
		TD conv		(64,1,1) (64,3,8) (256,1,1)	
				} 2	
Transition Block			(128,256)	(32,1,1)	(32,256)
Convolution Layer			(32,256)	(2,1,1)	(2,256)
Flatten			(2,256)	~	(512,)
Fully-connected Layer			(512,)	(129,)	(129,)

2.3. Transition block

With the number of connected layers increasing in dense block, the connected time channels are accumulated rapidly. To improve the model compactness, a transition block is inserted into two adjacent dense blocks to reduce the number of time channels. In our architecture, the transition block is a 1-D convolutional layer across frequency dimension and the time channels of the dense block output are compressed to 25% of their original value.

2.4. Network configurations

The 1-D frequency and time directional convolution layers are successively connected to construct an extension block to expand the number of time and frequency channels. The extension block enable the use of larger receptive field in the following dilated convolution structure. Once the extension block, dense block and transition block are stacked together, one convolutional layer with SELU activation is used to perform cross-channel pooling and dimension reduction. Finally, an output layer with linear activation is utilized for target clean speech spectra estimation. A detail configuration of the proposed architecture is given in Table 1. The dense blocks are shown in the brackets. The input and output size of the layers are specified as *timeChannels* × *frequenceChannels*. The layer hyper-parameters are given in (*outChannels*, *kernelSize*, *dilationRate*) formant.

3. EXPERIMENTS

Experiments are conducted using TIMIT database [25]. A total of 1000 sentences are selected for training and another 400 sentences excluded from the training speech are used to construct the test set. White Gaussian, factory1 and babble noises from the NOISEX-92 database [26] and railway noise from the Aurora2 database [27] are used as noise signals. The training sentences are added to the four noise types 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, each waveform is 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 normalized to zero mean and unit variance, and a reverse step is processed on the output. To measure the noise environment adaptation performance, the generalization ability test is evaluated for noise mismatch condition. The factory2 noise from the NOISEX-92 database and restaurant and street noise from the Aurora2 database, excluded from the training noise corpora, are used as the

Table 2: PESQ and STOI scores in noise match condition.

Metrics	PESQ					STOI				
Method	Noisy	DNN	LSTM	C-RNN	Proposed	Noisy	DNN	LSTM	C-RNN	Proposed
SNR 15	2.7007	2.8403	3.0342	3.1015	**3.2111**	0.9252	0.8664	0.8983	0.8992	**0.9280**
SNR 10	2.3713	2.7061	2.8671	2.9438	**2.9770**	0.8571	0.8405	0.8738	0.8768	**0.8964**
SNR 5	2.0415	2.4954	2.6180	**2.7070**	2.6924	0.7611	0.7914	0.8264	0.8349	**0.8433**
SNR 3	1.9110	2.3787	2.5088	**2.5843**	2.5654	0.7171	0.7641	0.8009	0.8106	**0.8149**
SNR 0	1.7254	2.1906	2.2610	**2.3876**	2.3656	0.6463	0.7094	0.7384	0.7622	**0.7622**
SNR -3	1.5503	1.9487	2.0325	**2.1520**	2.1291	0.5749	0.6403	0.6673	**0.7000**	0.6968
SNR -5	1.4532	1.7907	1.7767	**1.9812**	1.9777	0.5290	0.5883	0.5917	**0.6487**	0.6478
Avg.	1.9648	2.3358	2.4426	2.5510	**2.5598**	0.7158	0.7429	0.7710	0.7903	**0.7985**

Table 3: PESQ and STOI scores in noise mismatch condition

Metrics	PESQ					STOI				
Method	Noisy	DNN	LSTM	C-RNN	Proposed	Noisy	DNN	LSTM	C-RNN	Proposed
restaurant	2.0460	2.1762	2.3124	**2.3906**	2.3866	0.7120	0.7231	0.7567	0.7650	**0.7729**
street	2.2868	2.3934	2.6070	2.6982	**2.7371**	0.7932	0.7744	0.8150	0.8273	**0.8432**
factory2	2.2618	2.4708	2.6596	2.7219	**2.8019**	0.7910	0.7843	0.8213	0.8322	**0.8489**
Avg.	2.1982	2.3468	2.5263	2.6036	**2.6419**	0.7654	0.7606	0.7977	0.8082	**0.8217**

noise signals for generalization ability test. Two unseen SNR levels with -3 dB and 3 dB are also used for performance evaluation under noise match and noise mismatch conditions. We compare our proposed method with the following four baselines:

1. DNN [6]. DNN contains 3 hidden layers and each layer has 1024 hidden nodes.
2. LSTM [13]. LSTM contains 2 hidden layers, both of which has 1024 hidden units.
3. C-RNN [5]. C-RNN uses a 2-D convolution with 64 feature maps, kernel size (*T*, 16) and time-frequency stride (1, 8) to transform the input noisy spectra. The convolution output is then connected to a Bi-LSTM with 2 hidden layers of size 512.

The context frame of each method is all set as 11 (*T*=11). The efficient ADAM algorithm ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$) [28] is applied to train all the parameters of the DNN, LSTM, C-RNN and proposed architecture. All training vectors are subdivided into mini-batches with each containing 128 training vectors; the weights are updated after each mini-batch. The objective speech quality and intelligibility is evaluated via perceptual evaluation of speech quality (PESQ) [29] and Short-Time Objective Intelligibility (STOI) scores, respectively [30].

The average PESQ and STOI values of the noisy speech and enhanced speech by competition methods across four seen noise types are given in Table 2. Table 3 illustrates the average PESQ and STOI values across seven SNR levels (-5, -3, 0, 3, 5, 10, 15 dB) for unseen noises. Table 2 and Table 3 show that the proposed method produces consistently better PESQ and STOI performance than the 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. Table 4 shows the computational efficiency comparison between the baselines and proposed model and it reveals that our proposed method achieves a large boost on the computational efficiency.

Table 4: The computational efficiency comparisons of the baselines and proposed model.

Method	Number of parameters (Million)
DNN	3.68
LSTM	13.25
C-RNN	12.47
Proposed	0.75

4. CONCLUSIONS

A densely connected network with *T-F* dilated convolution was proposed for speech enhancement in this paper. The designed *T-F* dilated convolution block enlarge the receptive field and capture the contextual information in both temporal and frequency domains. The dense connectivity of each *T-F* convolution block ensures the maximum information flow between layers and alleviates the vanishing gradient problem. The 1-D convolution with the bottleneck structure is applied in the *T-F* convolution block to improve the computational efficiency. The performance evaluation shows that the introduced scheme produces consistently better enhancement performance than DNN and LSTM methods under seen and unseen noise conditions. Compared with C-RNN, the proposed method shows notable performance advantage under high SNR levels. The designed architecture improves the computation efficiency significantly, which shows the implementation potential to hardware.

12. REFERENCES

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

[2] 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.

[3] 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.

[4] 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, 1984.

[5] H. Zhao, S. Zarar, I. Tashev, and C. H. Lee, "Convolutional-Recurrent Neural Networks for Speech Enhancement," in ICASSP, 2018, pp. 2401-2405.

[6] 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.

[7] 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.

[8] D. W. Ke Tan, "A Convolutional Recurrent Neural Network for Real-Time SpeechEnhancement," in Interspeech, 2018, pp. 3229-3233.

[9] J. D. Tian Gao, Li-Rong Dai, Chin-Hui Lee, "Densely Connected Progressive Learning for LSTM-Based Speech Enhancement " in ICASSP, 2018, pp. 5054-5058.

[10] D. V. A. Babafemi O. Odelowo, "A Study of Training Targets for Deep Neural Network-Based Speech Enhancement Using Noise Prediction " in ICASSP, 2018, pp. 5409-5413.

[11] Y. Zhao, D. L. Wang, I. Merks, and T. Zhang, "DNN-based enhancement of noisy and reverberant speech," in ICASSP, 2016, pp. 6525-6529.

[12] 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.

[13] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. Schuller, *Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR*: Springer International Publishing, 2015.

[14] 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.

[15] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in Interspeech, 2017, pp. 3642-3647.

[16] K. Tan, J. Chen, and D. L. Wang, "Gated Residual Networks with Dilated Convolutions for Supervised Speech Separation," in ICASSP, 2018, pp. 21-26.

[17] S. R. Park, and J. Lee, "A Fully Convolutional Neural Network for Speech Enhancement," 2016, pp. 1993-1997.

[18] S. W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw Waveform-based Speech Enhancement by Fully Convolutional Networks," in APSIPA Annual Summit and Conference, 2017, pp. 6-12.

[19] D. Rethage, J. Pons, and X. Serra, "A Wavenet for Speech Denoising," in ICASSP 2018, pp. 5069 - 5073

[20] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech Enhancement Using Bayesian Wavenet," in INTERSPEECH, 2017, pp. 2013-2017.

[21] J. Chen, and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *Journal of the Acoustical Society of America,* vol. 141, no. 6, pp. 4705, 2017.

[22] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2261-2269.

[23] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," 2016.

[24] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-Normalizing Neural Networks," in NIPS, 2017, pp. 1-102.

[25] 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, 1993.

[26] 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.

[27] H. G. H. a. D. Pearce, "The AURORA experimental framework for the preformance evaluations of speech recognition systems under noisy conditions," in ISCA ITRW ASR, 2000, pp. 181–188.

[28] D. P. Kingma, and J. Ba, "Adam: A Method for Stochastic Optimization," in ICLR, 2015.

[29] 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 ICASSP, 2002, pp. 749-752 vol.2.

[30] 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.