INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multi-Scale TCN: Exploring Better Temporal DNN Model for Causal Speech Enhancement

Lu Zhang, Mingjiang Wang

Department of Electronics and Information Engineering,
Harbin Institute of Technology, Shenzhen, China, 518000
18B952047@stu.hit.edu.cn, mjwang@hit.edu.cn

Abstract

Capturing the temporal dependence of speech signals is of great importance for numerous speech related tasks. This paper proposes a more effective temporal modeling method for causal speech enhancement system. We design a forward stacked temporal convolutional network (TCN) model which exploits multi-scale temporal analysis in each residual block. This model incorporates a multi-scale dilated convolution to better track the target speech through its context information from past frames. Applying multi-target learning of log power spectrum (LPS) and ideal ratio mask (IRM) further improves model robustness, due to the complementarity among the tasks. Experimental results show that the proposed TCN model not only performs better speech reconstruction ability in terms of speech quality and speech intelligibility, but also has smaller model size than that of long short-term memory (LSTM) network and the gated recurrent units (GRU) network.

Index Terms: speech enhancement, multi-scale, temporal convolutional network, multi-objective learning.

1. Introduction

In the past few decades, there has been considerable interest in solving the noise interference of speech signals received in our real-life environments. Speech enhancement has been widely employed as a key front-end signal processing technique for various speech related products, such as hearing aids, smart phones and teleconferencing system. Despite its long research history, monaural speech enhancement is still a challenging subject in dealing with the complex and serious noise damage conditions.

Recently, deep neural networks (DNNs) has spurred the development of monaural speech enhancement, owed to their powerful modeling capacity on the relationship between corrupted and clean speech. Feedforward neural network (FNN) is the most widely used DNN model in the research field of speech enhancement. Many objective expressions, like ideal binary mask (IBM) [1], ideal ratio mask (IRM) [2] and log power spectrum (LPS) [3, 4], were proposed as training targets for supervised FNN-based speech denoising task. The human acoustic properties [5, 6] were also incorporated into the loss function of FNN model to achieve more comfortable enhanced speech. However, noise and speaker generalization problems exist in many FNN-based speech denoising methods, due to the characteristics of local frame modeling. The limited temporal windows of acoustic input features are not sufficient to decide the target speaker to focus on since the energy of target speech and noise fluctuates over time and the local signal-to-noise ratio (SNR) varies [7]. Although the usage of context information from past and future frames effectively improved the generalization problem in [4], it brought the non-causal problem for a real-time processing system.

Considering the temporal dependence of signals, recurrent neural networks (RNNs) have been utilized in [7, 8, 9, 10] to improve the generalization ability of DNN-based speech denoising models. Long short-term memory (LSTM) units and gated recurrent units (GRU) employed in those models help to capture longer context memory from past speech frames. It is found that the RNN-based models are more advantageous for low-latency speech enhancement system and it, without future frames, performs better than the FNN-based models with future frames. Furthermore, multi-objective learning strategy used for LSTM in [11, 12, 13] further improved the enhanced speech quality and intelligibility.

More recently, temporal convolutional network (TCN) model with causal dilated convolutions showed better memory superiority for sequential modeling tasks [14]. Inspired by this idea, we propose a novel multi-scale TCN model that stacks the input features forward into each residual block for speech enhancement. A multi-scale convolution method is proposed to enlarge and refine the receptive field of model. Specifically, the stacked input features are concatenated with the extracted features in each residual block to perform multi-scale analysis. Additionally, to fully utilize the underlying complementarity of different training targets, LPS and IRM are combined for multiple-target joint learning.

The rest of paper is organized as follows. The architecture of forward stacked multi-scale TCN model is introduced in Section 2. The details of the proposed multi-scale convolution method and multi-objective learning strategy are presented. In Section 3, experimental results of the proposed methods are provided. Finally, conclusions are drawn in Section 4.

2. Proposed Speech Denoising System

2.1. Forward stacked TCN model

For a standard feedforward neural network structure, it is hard to train a deep model with more than three hidden layers for speech denoising. Experiences with many visual recognition tasks tell us that the depth of representations is of central importance for achieving better model performance. In order to mitigate difficulties of training very deep models, a ResNet [15] structure was proposed to create some shortcuts for back-propagation by employing many skip-connected residual blocks (ResBlocks). Inspired by this, we propose a multi-scale temporal DNN framework for speech enhancement task, in which multiple residual blocks are sandwiched between two dense layers, as presented in Figure 1.

Previous research [16] has demonstrated that the widened architecture for residual blocks is conducive to improve the representation performance and speed of ResNet. Therefore, in

http://dx.doi.org/10.21437/Interspeech.2020-1104

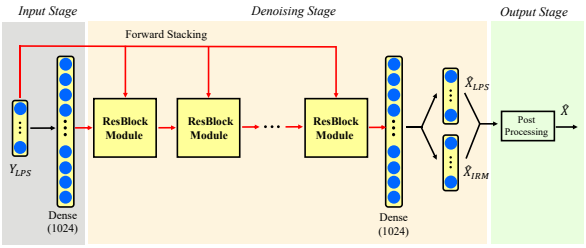


Figure 1: *Architecture of forward stacked TCN model for speech enhancement.*

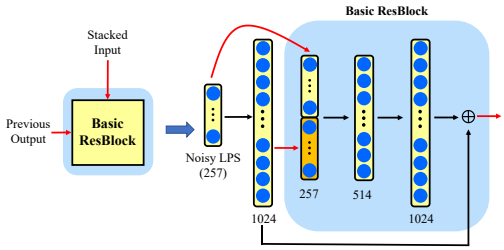


Figure 2: *Diagram of the basic ResBlock module with dilated convolution.*

our design, both dense layers are 1024 dimensions, aiming to extend the input feature to a high dimensional representation. Dilated convolutional layer is exploited in each residual block to capture the speaker's useful context information from past frames. The LPS features of noisy signals are extracted as the input of TCN model to learn clean LPS and IRM targets. In particular, the original noisy LPS features are stacked forward into each ResBlock to shorten the path of gradient propagation. Finally, a composite enhancement scheme is used, and the estimated LPS and IRM targets are combined to achieve better speech reconstruction in a post processing way.

2.2. Basic residual block

In our proposed TCN framework, ResBlock module plays an essential role in temporal modeling of signals. A basic residual block is firstly introduced to look back at a history of context for signal reconstruction. As shown in Figure 2, the basic ResBlock module is a three-layer bottleneck structure with skip connection. Only the middle convolutional layer uses dilated convolution, and the other two layers use standard 1-D convolutions. The kernel dimensions of three convolutional layers are 1×1024, 3×514, and 1×514, respectively. Their output channels are 257, 514, and 1024, respectively, to build up a widened bottleneck structure. Batch normalization [17], ReLU activation and dropout [18] are successively performed after each convolution operation. It should be noted that the first layer of each ResBlock module consists of two parts: a 1-D convolutional layer and a stacked original input feature. It means that the manually extracted features and the network automatically extracted features can be combined through the ResBlock module for a deeper representation.

Furthermore, using dilated convolutional layer enables the ResBlock module to represent a wider range of inputs:

$$F_d(t) = (Y * f_d) = \sum_{i=0}^{K-1} f_d(i) \bullet Y(t - d \cdot i) \qquad (1)$$

Where f_d and $F_d(t)$ represent the dilated convolution kernel and its output, respectively, t is the frame index, d is the dilation factor, K is the kernel size, and $Y(t - d \cdot i)$ accounts for the past frames for analysis. In order to avoid the gridding effect of dilated convolution [19], the choose of dilation factors should not be common divisors greater than 1. The skip-connected sum operation before the non-linear activation of last layer allows our TCN model to learn modifications to the identity mapping rather than the entire transformation.

2.3. Multi-Scale residual block

In our real life, due to the differences of word length and pronunciation characteristics (such as speech speed) of different people, the utterances always have the feature of temporal scale variation. Therefore, multi-scale methods [20, 21] have been investigated to remedy the problem of temporal scale variation. Using many branches with different receptive fields can improve the performance, but it increases the model size and processing burden.

As presented in Figure 3(a), we propose a simple yet efficient multi-scale ResBlock module to cope with the temporal scale variation. Unlike those branchy approaches with multiple parallel filters, the multi-scale of our proposed method refers to the multiple available receptive fields in one convolutional layer:

$$F_{md,b}(t) = (Y * f_{md,b}) = \sum_{i=0}^{K-1} f_{md,b}(i) \bullet \{\tilde{F}_{md,b-1}, Y_b\}(t - d \cdot i) \quad (2)$$

Where $f_{md,b}$ and $F_{md,b}(t)$ are the multi-scale dilated convolution kernel and its output of receptive sub-band b, respectively. $\tilde{F}_{md,b-1}$ are the output features of the previous sub-band. In this paper, 8 receptive sub-bands are divided in the first layer for multi-scale decomposition. As shown in Figure 3(b), multi-scale sub-band analysis is carried out in the middle layer from two directions. Output features of the previous sub-band are concatenated with the input of the next sub-band to perform dilated convolution operation. The number of output channels for each sub-band remains unchanged after each convolution. It should be noted that batch normalization, ReLU activation and dropout operations are successively performed after the convolution in each sub-band. This process repeats several times until all the divided sub-bands are analyzed. Finally, sum the obtained features of the two decomposition directions to integrate the results of the multi-scale analysis.

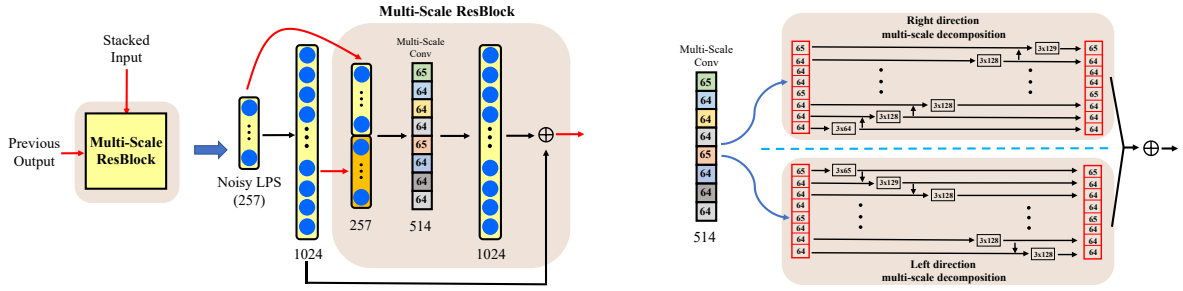
Like the basic ResBlock module, the multi-scale ResBlock also adopts the bottleneck structure to save model parameters. This introduced multi-scale method can linearly increase the equivalent receptive field in just one dilated convolution layer, while reducing the parameters of each residual block by 40%. Thus, it is more suitable for integration into the proposed TCN framework to obtain deeper feature representation and richer contextual history for speech enhancement.

2.4. Multi-Objective learning

To obtain better noise reduction ability, our idea is to jointly optimize the loss of LPS and IRM in our TCN framework:

$$Loss_{MT} = \frac{1}{T \cdot M} \sum_{t=1}^{T} \sum_{k=1}^{M} [(\hat{X}_{LPS}(k,t) - X_{LPS}(k,t))^2 + (\hat{X}_{IRM}(k,t) - X_{IRM}(k,t))^2] \qquad (3)$$

Where $\hat{X}_{LPS}(k,t)$ and $\hat{X}_{IRM}(k,t)$ are the enhanced LPS and the estimated IRM, respectively. Correspondingly, $X_{LPS}(k,t)$



(a) (b)

Figure 3: *(a) Diagram of the multi-scale ResBlock module with dilated convolution. (b) Principles of the multi-scale decomposition in left and right direction.*

and $X_{IRM}(k,t)$ are their ideal targets. M and T represent the feature length and batch size, respectively. Joint optimization of different objectives is equivalent to incorporating multiple regularization terms for the training of TCN model. Therefore, in the output stage of model test, the estimated LPS features and IRM enhanced speech can be averaged in magnitude spectrum domain as the final output:

$$\hat{X} = \frac{1}{2}(\sqrt{\exp(\hat{X}_{LPS})} + \exp(Y_{LPS}) \bullet \hat{X}_{IRM}) \quad (4)$$

The result is then used with the noisy phase to reconstruct the waveform of the enhanced speech. Although the incorporation of dual targets leads to a slight increase in model parameters, the composite enhancement of both outputs can achieve their complementary advantages, thus better speech reconstruction.

3. Experimental results

3.1. Experimental setups

The experiments below were conducted on TIMIT database [22], which contains 4620 training utterances and 1680 test utterances. Four noise recordings (Babble, Factory1, Destroyer engine, and Destroyer operation noises) from the NOISEX-92 database [23] were selected to generate noisy database for model training. Each noise recording is about 4 minutes long. To construct the noisy training set, we used random cuts from the first 60% of each noise to mix with all training utterances of TIMIT database, and the mixed SNR follows the uniform distribution in the range of -5 to 15. Likewise, the middle 20% of each noise were used to mix with 280 utterances from the TIMIT test set to construct the validation set for model training. To evaluate the noise and speaker generalization ability of DNN models, 320 unseen utterances from the TIMIT test set were mixed with the last 20% of each noise to construct the test set of seen noise cases. Besides, two new noise types from NOISEX-92, namely *Pink* and *Factory2*, were used as the unseen noise cases for test. The short-time objective intelligibility (STOI) [24] and perceptual evaluation of speech quality (PESQ) [25] were adopted as two metrics to evaluate the speech denoising performance.

As for signal analysis, all the speech and noise signals were resampled to 16 kHz, and the frame length was set to 512 samples with a frame shift of 256. Thus, 257-dimensional LPS features were fed into DNN models for training. The Adam algorithm [26] was used to optimize the model parameters with a learning rate of 0.001 in every mini-batch. The dropout rate in our proposed TCN models was set to 0.2. All the input LPS features were normalized to zero mean and unit variance for model training and test. Noted that the test noisy speech was normalized by the global mean and variance of all training data to ensure the causality of our system.

3.2. Comparison between different temporal DNN models

To evaluate the speech denoising effect of the proposed multi-scale TCN models in causal speech enhancement task, several classic DNN structures were compared experimentally. The evaluated PESQ and STOI results are presented in Table 1 and Table 2, respectively. We used an FNN model with 3 hidden layers of size 2048 as a baseline system without contextual frames (denoted as "FNN-SE"). "LSTM-SE" and "GRU-SE" are two widely used temporal models for speech enhancement with memory mechanism. Both RNN-based models contain 3 hidden layers with 1024 neurons per layer. The proposed TCN models with the basic ResBlock and the multi-scale ResBlock are denoted as "TCN-SE" and "MSTCN-SE-1", respectively. Unlike the above models that only learn a single LPS target, "MSTCN-SE-2" represents the multi-objective learning model of LPS and IRM. All the proposed TCN models stacked 5 ResBlocks, and the dilated rate increased gradually, which was 1, 2, 5, 7, 11, respectively.

From Table 1 and 2, it is observed that the temporal DNN methods utilizing the context information of current and past frames show better speech denoising effect than that of FNN-SE modeled only in current frame. That is, the temporal models have better speaker and noise generalization capability

Table 1: *Averaged PESQ results obtained for noisy and enhanced speech in seen and unseen noise cases*

PESQ Results		**Input SNRs (dB)**				
Noises	Methods	-5dB	0dB	5dB	10dB	15dB
Seen	Noisy	1.29	1.63	1.99	2.35	2.71
	FNN-SE	1.66	2.17	2.60	2.95	3.22
	LSTM-SE	1.74	2.24	2.63	2.92	3.14
	GRU-SE	1.75	2.24	2.62	2.92	3.14
	TCN-SE	2.02	2.44	2.73	2.95	3.10
	MSTCN-SE-1	2.03	**2.50**	**2.84**	3.10	3.30
	MSTCN-SE-2	**2.06**	2.49	**2.84**	**3.14**	**3.40**
Unseen	Noisy	1.32	1.68	2.06	2.45	2.81
	FNN-SE	1.53	2.05	2.50	2.87	3.16
	LSTM-SE	1.69	2.19	2.61	2.92	3.15
	GRU-SE	1.78	2.25	2.63	2.93	3.16
	TCN-SE	2.03	2.44	2.73	2.94	3.10
	MSTCN-SE-1	**2.12**	**2.55**	2.87	3.12	3.31
	MSTCN-SE-2	2.08	2.53	**2.88**	**3.18**	**3.44**

Table 2: *Averaged STOI results obtained for noisy and enhanced speech in seen and unseen noise cases*

STOI Results		Input SNRs (dB)				
Noises	Methods	-5dB	0dB	5dB	10dB	15dB
Seen	Noisy	0.547	0.663	0.772	0.861	0.924
	FNN-SE	0.627	0.756	0.841	0.889	0.915
	LSTM-SE	0.629	0.764	0.846	0.892	0.917
	GRU-SE	0.645	0.767	0.845	0.891	0.915
	TCN-SE	0.704	0.811	0.868	0.900	0.918
	MSTCN-SE-1	0.713	0.821	0.880	0.915	0.935
	MSTCN-SE-2	**0.720**	**0.827**	**0.890**	**0.931**	**0.958**
Unseen	Noisy	0.592	0.704	0.807	0.887	0.941
	FNN-SE	0.596	0.733	0.828	0.885	0.913
	LSTM-SE	0.607	0.749	0.843	0.894	0.919
	GRU-SE	0.638	0.764	0.845	0.892	0.916
	TCN-SE	0.700	0.810	0.870	0.902	0.920
	MSTCN-SE-1	0.727	0.827	0.885	0.918	0.938
	MSTCN-SE-2	**0.729**	**0.836**	**0.898**	**0.937**	**0.961**

than FNN-SE in speech enhancement task. Among the above temporal models, the three TCN models proposed in this paper achieve better quality and intelligibility of enhanced speech in both seen and unseen noise cases. In contrast to the LSTM-SE and GRU-SE models, our basic TCN-SE model obtains a notable improvement of PESQ and STOI at low SNR cases of -5~5 dB. The presented multi-scale TCN models refine the temporal analysis of speech signals, which is beneficial to recover more details of speech spectrum. The spectral filtering operation of IRM further compensates the speech distortion problem of LPS at high SNR cases (10~15 dB). Therefore, combining the benefits of IRM and LPS targets enables the MSTCN-SE-2 model to achieve the best PESQ and STOI results.

Moreover, the model sizes of the above DNN models are presented in Table 3. The number of trainable parameters of MSTCN-SE-1 is less, only 7.4 million, while FNN-SE, LSTM-SE and GRU-SE are 9.5 million, 22.3 million and 16.8 million, respectively. Multi-scale convolution contributes the improvement of computational efficiency and noise reduction ability. Although the multi-objective learning strategy slightly increases the trainable parameters of the MSTCN-SE-2 model, it guarantees better enhanced speech quality and intelligibility.

Table 3: *Model size of different DNN models*

Model Size (million)	FNN-SE	LSTM-SE	GRU-SE	TCN-SE	MSTCN-SE-1	MSTCN-SE-2
	9.5	22.3	16.8	9.8	7.4	7.7

3.3. Comparison with previous multi-objective methods

In this section, we compared the evaluation results of PESQ and STOI between our MSTCN-SE-2 model and two RNN-based multi-target learning methods for speech enhancement. The results are presented in Figure 4 and 5. "LSTM-SE-MT" represent the LSTM-based learning method of IRM and LPS targets [11], and "LSTM-SE-PL" is the densely connected LSTM progressive learning model with 5 LPS targets [12].

Figure 4 and 5 illustrate that the proposed MSTCN-SE-2 consistently outperforms the other two LSTM models at all SNR cases. This performance superiority is more significant at low SNR cases (-5 and 0 dB). The LSTM network is more

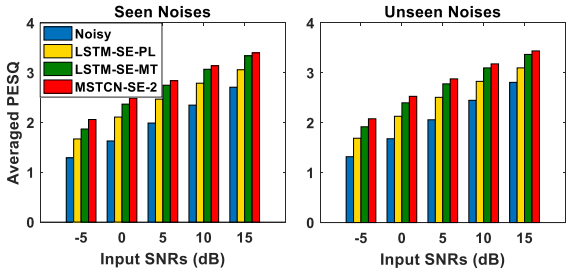


Figure 4: *Averaged PESQ results obtained from different multi-objective methods in seen and unseen noise cases*

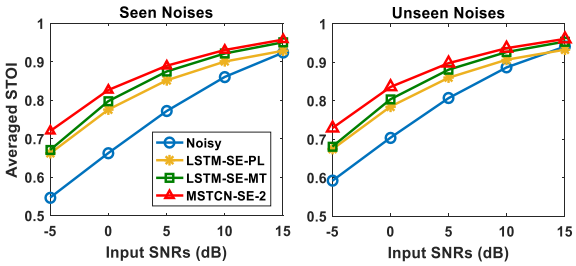


Figure 5: *Averaged STOI results obtained from different multi-objective methods in seen and unseen noise cases*

susceptible to the starting point of input, and its long-term dependence is easy to introduce more useless information. In contrast, the most important local information is considered in each forward stacked ResBlock module of MSTCN-SE-2 for speech signal analysis. In terms of model size, the trainable parameters of the LSTM-SE-MT and LSTM-SE-PL models are 14.2 million and 38.2 million, respectively, which are much larger than the proposed MSTCN-SE-2. It indicates that the proposed multi-scale dilated convolution contributes to more excellent temporal modeling ability for speech signals than the classical LSTM units, while saving more parameters.

4. Conclusions

This paper presents a more efficient multi-scale TCN model for monaural speech enhancement. A novel multi-scale dilated convolution method is proposed to enlarge the receptive field of ResBlock at a more granular level. The strategy of stacking input features and skip connection in each ResBlock enables us to train a deeper model for feature representation. Owe to these advantages on analyzing the contextual information of speakers, the proposed TCN methods exhibit better denoising effect and stronger model generalization than the other DNN temporal modeling methods. In addition, the presented multi-objective learning architecture fully utilizes the advantages of LPS and IRM, and improves the robustness of the model under various noise damage levels.

5. Acknowledgements

This work was supported in part by the Basic Research Program under Grants No. JCYJ20170412151226061 and No. JCYJ20180507182241622 funded by Shenzhen government.

6. References

[1] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[2] 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.

[3] 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.

[4] 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 and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[5] W. Han, X.-W. Zhang, M. Sun, "Perceptual improvement of deep neural networks for monaural speech enhancement," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1-5.

[6] Y. Zhao, B. Xu, R. Giri, "Perceptually guided speech enhancement using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5074-5078.

[7] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2016, pp. 3314-3318.

[8] J. M. Valin, "A Hybrid DSP/deep learning approach to real-time full-band speech enhancement," in *International Workshop on Multimedia Signal Processing (MMSP)*, 2018, pp. 1-5.

[9] Y. Tu, I. Tashev, S. Zarar and C.-H. Lee, "A Hybrid Approach to Combining Conventional and Deep Learning Techniques for Single-Channel Speech Enhancement and Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2531-2535.

[10] Y. Luo, N. Mesgarani, "TasNet: time-domain audio separation network for real-time, single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 696-700.

[11] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Hands-free Speech Communications and Microphone Arrays (HSCMA)*, 2017, pp. 136-140.

[12] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Densely connected progressive learning for LSTM-based speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5054-5058.

[13] J. Lee, J. Skoglund, T. Shabestary and H.-G. Kang, "Phase-sensitive joint learning algorithms for deep learning-based speech enhancement," *IEEE Signal Processing Letters*, vol 25, no. 8, pp. 1276-1280, 2018.

[14] S. J. Bai, J. Kolter and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv: 1803.01271*, 2018.

[15] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.

[16] S. Zagoruyko, N. Komodakis, "Wide Residual Networks," in *British Machine Vision Conference (BMVC)*, 2016, pp. 1-12.

[17] S. Loffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Learning Representations (ICLR)*, 2015, pp. 448–456.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol 15, pp. 1929–1958, 2014.

[19] P. Wang, P. Chen, and Y. Yuan, "Understanding convolution for semantic segmentation," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 1451-1460.

[20] N. Takahashi, and Y. Mitsufuji, "Multi-Scale multi-band densenets for audio source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 21-25.

[21] Z. Shi, H. Lin, L. Liu, R. Liu, J. Han, and A. Shi, "End-to-End monaural speech separation with multi-scale dynamic weighted gated dilated convolutional pyramid network," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 4614-4618.

[22] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database NIST Tech Report," 1988.

[23] A. Varga, and H. J. 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, pp. 247–251, 1993.

[24] 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.

[25] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001, pp. 749–752.

[26] D. Kingma, and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015, pp.1-13.