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

Bidirectional LSTM Network with Ordered Neurons for Speech Enhancement

*Xiaoqi Li, Yaxing Li, Yuanjie Dong, Shan Xu, Zhihui Zhang, Dan Wang, Shengwu Xiong**

School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China
{xli, xiongsw}@whut.edu.cn, whhit173@hotmail.com

Abstract

Speech enhancement aims to reduce the noise and improve the quality and intelligibility of noisy speech. Long short-term memory (LSTM) network frameworks have achieved great success on many speech enhancement applications. In this paper, the ordered neurons long short-term memory (ON-LSTM) network with a new inductive bias to differen-tial the long/short-term information in each neuron is pro-posed for speech enhancement. Comparing the low-ranking neurons with short-term or local information, the high-ranking neurons which contain the long-term or global in-formation always update less frequently for a wide range of influence. Thus, the ON-LSTM can automatically learn the clean speech information from noisy input and show better expressive ability. We also propose a rearrangement concat-enation rule to connect the ON-LSTM outputs of forward and backward layers to construct the bidirectional ON-LSTM (Bi-ONLSTM) for further performance improvement. The experimental results reveal that the proposed ON-LSTM schemes produce better enhancement performance than the vanilla LSTM baseline. And visualization result shows that our proposed model can effectively capture clean speech components from noisy inputs.

Index Terms: order neurons, long-short term memory network, speech enhancement

1. Introduction

In the past few years, deep learning based speech enhancement approaches have achieved very promising performance. Based on a definition of the learning targets, the data driven approaches are categorized as spectral mapping [1, 2] and time-frequency masking [3, 4, 5]. Mapping approaches with regression-based deep neural network (DNN) predict the magnitude or power spectra of clean speech from the noisy speech [1, 2]. Masking-based targets predict the mask to produce a spectrogram of clean speech when applied to noisy input audio, such as the ideal ratio mask (IRM) [3], the phase-sensitive mask (PSM) [4] and complex-valued ratio mask (cRM) [5]. Except the learning targets, the supervised speech enhancement methods also are investigated from the aspects of network structure. Generative adversarial networks (GANs) [6, 7], variational autoencoders (VAEs) [8, 9] and WaveNet [10, 11] for speech denoising have been proposed. The network structures are evolved from fully-connected layer network [1, 2] to recurrent neural network (RNN) [12, 13, 14], convolution neural network (CNN) [15, 16, 17, 18] and their combinations [19, 20, 21]. RNN is highly effective at the task of speech processing and LSTM [22] alleviates the gradient vanishing or exploding issues in standard RNNs by using input, output and forget gates. LSTM has been investigated for speech enhancement [12, 13] to capture the temporal dependences of speech signal and significantly outperforms the DNN with feedforward structure.

Recently, ordered neurons [23], a new inductive bias for recurrent neural networks, have been proposed and achieved good performance at the task of language modeling. This inductive bias promotes differentiation of long/short-term information stored inside each neuron: long-term information which keeping for a large number of steps will be stored in high-ranking neurons, while low-ranking neurons storing short-term information that can be rapidly forgotten [23]. It is intuitive to use hierarchical information in language sequences. The impact of a new input on the sequence is often small, and the main information from hidden layer (such as global information) is still affected by the historical state. The new input only changes the high-level information when the global information of the sequence changes and the influence of historical information becomes smaller. In other words, high-ranking neurons, which will last anywhere from several time steps to the entire sentence, always update less frequently than the others while the low-level information changes continuously with the current input.

The speech signal also shows some kind of hierarchical information analogous to the natural language. The high-level information includes the speaker information, speech emotion and so on, while the low-level speech information includes current phoneme, etc. In this work, a speech enhancement framework based on ON-LSTM is proposed. To the best of our knowledge, this is the first work that ON-LSTM is applied to speech-related task. The ON-LSTM can automatically learn the hierarchical information of speech signal, mainly about the existence of speech, and shows better expressive ability. We also proposed a rearrangement concatenation rule to connect the ON-LSTM outputs of forward and backward layers to construct the Bi-ONLSTM structure for speech enhancement. The experimental results show that the proposed ON-LSTM and Bi-ONLSTM schemes produce satisfactory enhancement performance comparing the other baselines.

The remainder of this paper is organized as follows. In Section 2, we first present our newly proposed architecture for speech enhancement. We also explore a rearrangement concatenation rule in the Bi-ONLSTM structure. Then we show a series of experiments to verify the effectiveness of our methods in Section 3. Finally, we conclude our work in Section 4.

2. ON-LSTM Based Speech Enhancement

In this section, we introduce our ON-LSTM architecture for speech enhancement, including the differences between LSTM and ON-LSTM, the structure of Bi-ONLSTM and the rearrangement concatenation rule for the outputs of the forward and backward layers.

* Corresponding author

This work was in part supported by the National Natural Science Foundation of China (Grant No. 61702386), the National Key Research and Development Program of China (Grant No. 2017YFB1402203), the Defense Industrial Technology Development Program (Grant No. JCKY2018110C165), Major Technological Innovation Projects in Hubei Province (Grant No. 2019AAA024). This work was also supported by "the Fundamental Research Funds for the Central Universities".

http://dx.doi.org/10.21437/Interspeech.2020-2245

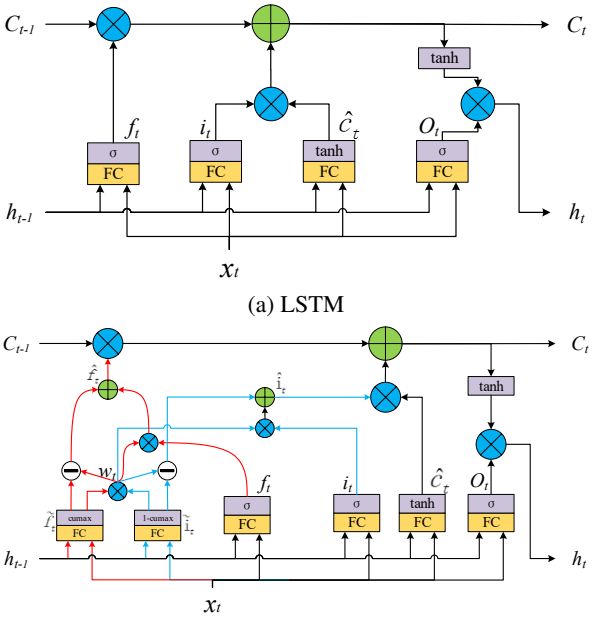


Figure 1: *The architecture of LSTM and ON-LSTM. "FC" stands for a fully-connected layer. The red lines and blue lines indicate the process of calculating $\hat{f}_t$ and $\hat{i}_t$.*

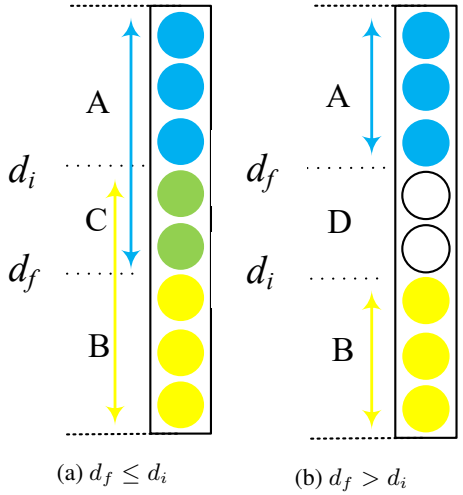


Figure 2: *Division of the segments in hidden neurons. Updating rules in different parts: A, retain only the original historical information; B, retain only the current input information; C, update with the vanilla LSTM; D, no information input.*

2.1. ON-LSTM Structure

The architecture of LSTM and ON-LSTM are illustrated in Figure 2. Before getting into the details of the ON-LSTM layer in our architecture, we first revisit the standard LSTM.

Since the gates in the LSTM act independently on each neuron, the neurons are "disordered" and it may be difficult to discern a hierarchy of information between the neurons. In ON-LSTM, high-ranking neurons contain long-term or global information that will last anywhere from several time steps to the entire sequence for a wide range of influence. Low-ranking neurons encode short-term or local information that only last one or a few time steps to represent smaller constituents. Therefore, it is necessary to make the neuron whose index of cell states C_t is smaller indicates the lower level information, and the neuron with the larger index represents the higher level information. So the difference between LSTM and ON-LSTM will be the update mechanism from $\hat{C}_t$ and C_{t-1} to C_t.

Two integers d_f and d_i are used to divide the hidden layers into two segments, in which different update rules are applied to differentiate the long/short-term information. The information stored in the first d_f neurons of the previous cell state C_{t-1} that represents the end of a high-level constituent in the hidden vector is erased completely. The information provided by the current input x_t is written into the first d_i neurons of the current cell state C_t. A large d_i indicates that the current information contains the long-term information and conversely a small d_i means that the current input x_t just provides local information.

The two segments are overlapped if the integer d_f is less than or equal to d_i. As shown in Figure 2. (a), the update rules with the standard LSTM are applied to the neurons of the overlapped region C. The original historical information and the current input information are retained for the region A (above index d_i) and region B (below index d_f), respectively. As shown in Figure 2. (b), the two segments are not overlapped since the integer d_f is larger than d_i. The historical information and current input information are retained for the region A and region B, respectively, and the neurons of the remaining region D is set to zero with no information input.

We use a binary gate $g = (0, 0, ..., 0, 1, ..., 1)$ to split the cell state into 0-segment and 1-segment, in which the corresponding neurons are updated with more and less frequencies, respectively. The master forget gate $\tilde{f}_t$ and the master input gate $\tilde{i}_t$ are newly introduced to control the erasing and the writing behaviors, respectively. The split point of $\tilde{f}_t = (0, ..., 0, 1, ..., 1)$ and $\tilde{i}_t = (1, ..., 1, 0, ..., 0)$ are d_f and d_i, respectively. Since such binary gates are not differentiable, the backpropagation algorithm can not be used to train the neural networks directly. A new activation function in [28] given below is introduced for the backpropagation training of the master gates:

$$g = cumax(\cdot) = cumsum(softmax(\cdot)) \tag{1}$$

$$\tilde{f}_t = cumax(W_{\tilde{f}} h_{t-1}) + U_{\tilde{f}} x_t + b_{\tilde{f}}) \tag{2}$$

$$\tilde{i}_t = 1 - cumax(W_{\tilde{i}} h_{t-1}) + U_{\tilde{i}} x_t + b_{\tilde{i}}) \tag{3}$$

where W, U are the weight matrices and b is the bias vectors.

The $cumax(\cdot)$ obtained by taking a cumulative sum of the softmax is used to approximate the gate. The values in the master gate $\tilde{f}_t$ and $\tilde{i}_t$ are monotonically increasing from 0 to 1 and decreasing from 1 to 0, respectively, and the different values differentiate the update frequencies.

Based on this activation function, ON-LSTM introduces novel ordered neuron rules to update cell state:

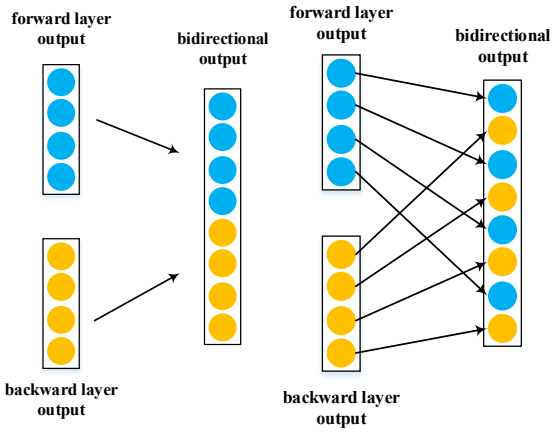
$$w_t = \tilde{f}_t \circ \tilde{i}_t \tag{4}$$

$$\hat{f}_t = f_t \circ w_t + (\tilde{f}_t - w_t) \tag{5}$$

$$\hat{i}_t = i_t \circ w_t + (\tilde{i}_t - w_t) \tag{6}$$

$$C_t = \hat{i}_t \circ \hat{C}_t + \hat{f}_t \circ C_{t-1} \tag{7}$$

In practice, a hyper-parameter C namely the chunk size is introduced to reduce the number of parameters for $\tilde{f}_t$ and $\tilde{i}_t$. We set $\tilde{f}_t$ and $\tilde{i}_t$ to be $D = N/C$ dimensional vectors, where N



(a) original concatenation rule (b) rearrangement concatenation rule

Figure 3: *The rearrangement concatenation rule for bidirectional output.*

represents the number of hidden layer neurons. We repeat each dimension C times after the calculation of $\tilde{f}_t$ and $\tilde{i}_t$.Therefore, ON-LSTM only increases the computation complexity slightly compared to the vanilla LSTM.

We propose a speech enhancement method based on ON-LSTM. We hope that the network will learn to determine the existence and duration of speech from noisy input to better extract clean speech components. If the current inputs start to contain clean speech segments, the update frequency of the input gate should be higher and the forget gate update with less frequency. When the input starts to contain only noise fragments, that is, the speech fragments disappear, then the update frequency of the forget gate should be higher and historical speech information will be erased mostly. But some higher-level information like speaker information will last until the end of the utterance so that not disturb the enhanced speech quality.

2.2. Bi-ONLSTM Structure and Rearrangement Concatenation

The unidirectional RNNs process the inputs in temporal order and their outputs tend to be mostly based on the previous context. There are ways such as adding a delay between the outputs and the targets to introduce future context, but these usually do not make full use of backwards dependencies [24]. In speech enhancement task, more acoustic future information could reduce the discontinuity of the estimated clean speech signals to obtain a better listening quality [2]. A better performance can be achieved if we replace the RNN with bidirectional RNN (Bi-RNN). In this paper, we develop the Bi-ONLSTM on the basis of ON-LSTM for speech enhancement.

The concatenated output of Bi-ONLSTM in each time step should be in a specific order as in ONLSTM. Generally, the output of the forward and backward LSTM are concatenated vertically as the output of Bi-LSTM, as shown in the Figure 3.(a). The vertical concatenation will break the neurons orderliness, which results in the performance degradation of the network and even the ambiguous of output vector. Therefore, we propose a new namely rearrangement concatenation rule to connect the forward and backward outputs in Bi-ONLSTM. As shown in the Figure 3.(b), we take a single neuron or a chunk from the forward and backward outputs as a unit and concatenate them alternately as the final output of Bi-ONLSTM. This concatenation rule will ensure the orderliness of the bidirectional output.

2.3. Network Architecture

In order to fairly compare the performance of LSTM and ON-LSTM for speech enhancement, all models contain 3 hidden layers. The last time step outputs of last layer in LSTMs are feed into one fully-connected layer with 129 nodes. We get the outputs of last bidirectional layer by concatenating the last hidden state of forward layer and first hidden state of backward layer. The final estimated clean speech features are obtained by a fully-connected layer with sigmoid activation and IRM is used as the training target. Dropout regularization [25] is used to reduce overfitting and dropout rates are 0.2.

3. Experiments

3.1. Experimental Settings

In our experiments, the TIMIT corpus [26] is used to prepare the training and test sets. A total of 1000 sentences from the training set of the TIMIT database are selected for training and another 400 sentences excluded from the training speech are used to construct the test set. Babble, F16, White Gaussian and Factory1 noises from the NOISEX-92 database [27] and another two noise types, namely office and restaurant, are used as noise signals. The training sentences are added to the six noise types to generate a set of artificially noisy utterances with four signal-to-noise-ratios (SNRs) levels (from -5 to 10 dB with 5 dB increments) to form 24000 training utterances. For the signal analysis, the original raw waveforms are first down sampled to 8 kHz and a 256-point Hamming window is applied with a 50% overlap. We use short-time Fourier transform (STFT) to extract the spectrogram from each utterance and the speech spectra are then represented by the 129 dimensional log power spectrum (LPS) features. The input features are standardized to have zero mean and unit variance. After the feature transformation steps are completed, 10% of the training features are assigned as the validation set. Except the four SNRs in the training set, two unseen SNR levels with -3 dB and 3 dB are also used for performance evaluation. In inference, the estimated spectrum and the phase from noisy speech will be converted to the waveform via inverse STFT.

To prevent the artificial distortion, so-called musical noise, in IRM processing, the flooring with a lower threshold value 0.05 is applied to the estimated mask before T-F mask processing [28, 29].

The mean squared error (MSE) is used as the objective function. The input context window of each method is set as 11 ($T = 11$) and it spans from past 5 frames to future 5 frames. All networks are trained using Adam optimizer [30] and the initial learning rate is set to 0.001 with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1.0e^{-8}$. Total number of epoch is set to 100 where the batch size is 128. When the validation loss doesn't decrease for more than 5 epochs, the training will be terminated in advance to reduce overfitting. For all the experiments, we select the trained model which produces the best MSE performance on the validation set. The objective speech quality and intelligibility are evaluated via perceptual evaluation of speech quality (PESQ) [31] and Short-Time Objective Intelligibility (STOI) scores, respectively [32].

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 denoising methods:

Table 1: *The average PESQ and STOI comparison of different models.*

methods	nodes (chunk size)	params (million)	PESQ	STOI
Noisy	-	-	1.8614	0.6822
LSTM	256	1.48	2.5147	0.8062
LSTM	512	5.58	2.5396	0.8095
Bi-LSTM	256(16)	4.01	2.5625	0.8160
ON-LSTM	256(64)	1.49	2.5988	0.8229
ON-LSTM	256(16)	1.52	2.6066	0.8232
ON-LSTM	256(4)	1.66	2.5879	0.8223
ON-LSTM	512(16)	5.67	2.6293	0.8295
Bi-ONLSTM	256(16)	4.13	**2.6608**	**0.8301**

Table 2: *The performance of different models in terms of PESQ and STOI.*

method	PESQ	STOI
Noisy	1.8614	0.6822
CRNN	2.5825	0.8160
DenseNet	2.5435	0.7983
Non-LocalNet	2.5824	0.8070
Bi-ONLSTM	**2.6608**	**0.8301**

CRNN: CRNN combines both convolutional and recurrent neural networks. Two layers of bidirectional LSTMs with 256 nodes follow 6 convolutional layer with kernel size of (1, 3).

DenseNet [22]: DenseNet contains 2 dense blocks with time-frequency (T-F) dilated convolution for speech enhancement.

Non-local Net [23]: A lightweight convolutional neural network with non-local module, which is capable of capturing the long-range dependencies in the frequency domain.

3.2. Results Comparison

Table shows the average PESQ and STOI values of the noisy speech and enhanced speech by LSTM and ON-LSTM. The ON-LSTM method produces better PESQ and STOI performance than the vanilla LSTM approaches regardless of the selected training target. The bidirectional models outperform the corresponding unidirectional models and Bi-ONLSTM achieves the best performance. Comparing the LSTM with 512 units and Bi-LSTM with 256 units, we find that the performance improvement of bidirectional model is not due to the increase of trainable parameters in network. It is also clear that the ON-LSTM whose chunk size is 16 could achieve a better performance in our experiment.

As illustrated in Tables 2, our proposed method produces better performance compared with three another models. Bi-ONLSTM improves PESQ by 0.08 and 0.08, and improves STOI by 2.3% and 1.4% over Non-Local Net and CRNN, respectively.

3.3. Analysis of Order Neurons

To explicitly under the role of order neurons in our task, we define the distance value Dis_t as:

$$\mathrm{Dis}_t = 1 - \frac{1}{D}\sum_{k=1}^{D} \tilde{f}_{tk} \tag{8}$$

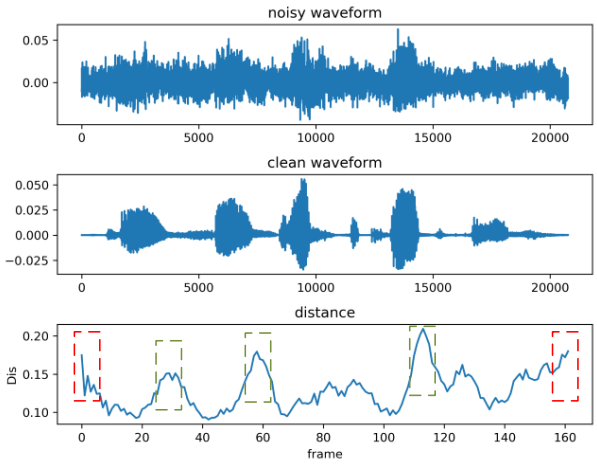


Figure 4: *Visualization of the waveform of noisy speech, clean speech and distance value from ON-LSTM. Cell size of model is 256 and chunk size is 16.*

where $\tilde{f}_{tk}$ denotes the k-th chunk value of master forget gate $\tilde{f}$ in t-th frames. The larger the value of Dis, the more historical speech information is updated. The last layer of ON-LSTM is used for evaluation. In our experiment, the context window size is 11 and we average Dis of the same frame in different contextual inputs.

Figure 4 shows a visualization example of the distance value between different frames on a testing noisy speech. We can clearly see that the larger values of Dis appear at the beginning and end of the utterance. In addition, at the end of each envelope, the distance value is also relatively large, especially in the portion marked with the green rectangular boxes. This observation shows that the proposed ON-LSTM network learn to discriminate the existence and duration of speech to a certain extent, and captures clean speech components from noisy inputs.

4. Conclusions

The ON-LSTM network is introduced for speech enhancement in this paper. The ON-LSTM can automatically learn the hierarchical structure information from noisy input, mainly about the duration of speech, and show better expressive ability due to its specifically ordered hidden neurons. The evaluation results reveal that the proposed framework provided consistently better enhancement performance than the vanilla LSTM regarding the PESQ and STOI metrics. In addition, we implement the Bi-ONLSTM and propose a new concatenation rule to connect the ON-LSTM outputs of forward and backward layers. The experimental results also show that the Bi-ONLSTM architecture outperforms the other speech enhancement methods. We believe that ON-LSTM will play an outstanding role in many other speech-related tasks and replace the ordinary LSTM in many hybrid network architectures for better performance.

5. References

[1] 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.

[2] 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, 2014.

[3] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7092–7096.

[4] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

[5] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[6] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *Proc. Interspeech 2017*, pp. 3642–3646, 2017.

[7] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 106–110.

[8] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2018, pp. 1–6.

[9] S. Leglaive, U. Şimşekli, A. Liutkus, L. Girin, and R. Horaud, "Speech enhancement with variational autoencoders and alpha-stable distributions," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 541–545.

[10] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[11] 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.

[12] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[13] 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.

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

[15] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," *Proc. Interspeech 2017*, pp. 1993–1997, 2017.

[16] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for supervised speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 21–25.

[17] Y. Li, X. Li, Y. Dong, M. Li, S. Xu, and S. Xiong, "Densely connected network with time-frequency dilated convolution for speech enhancement," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6860–6864.

[18] X. Li, Y. Li, M. Li, S. Xu, Y. Dong, X. Sun, and S. Xiong, "A convolutional neural network with non-local module for speech enhancement." in *INTERSPEECH*, 2019, pp. 1796–1800.

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

[20] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Interspeech*, 2018, pp. 3229–3233.

[21] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6865–6869.

[22] S. Hochreiter, J. urgen Schmidhuber, and C. Elvezia, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] Y. Shen, S. Tan, A. Sordoni, and A. Courville, "Ordered neurons: Integrating tree structures into recurrent neural networks," in *International Conference on Learning Representations*, 2019.

[24] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm networks," in *2005 IEEE International Joint Conference on Neural Networks*, vol. 4. IEEE, 2005, pp. 2047–2052.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[26] 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," *STIN*, vol. 93, p. 27403, 1993.

[27] 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, no. 3, pp. 247–251, 1993.

[28] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "Dnn-based source enhancement to increase objective sound quality assessment score," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1780–1792, 2018.

[29] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*, 2019, pp. 2031–2041.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[31] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[32] 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.