Study on convolutional recurrent neural networks for speech enhancement in fiber-optic microphones

Wenli Peng, Shenglai Zhen*, Xin Chen, Qianjing Xiong and Benli Yu

Information materials and intelligent perception Laboratory, Anhui University, Hefei, Anhui, 230601, China

*Corresponding author’s e-mail: slzhen@ahu.edu.cn

Abstract. In this paper, several improved convolutional recurrent networks (CRN) are proposed, which can enhance the speech with non-additive distortion captured by fiber-optic microphones. Our preliminary study shows that the original CRN structure based on amplitude spectrum estimation is seriously distorted due to the loss of phase information. Therefore, we transform the network to run in time domain and gain 0.42 improvement on PESQ and 0.03 improvement on STOI. In addition, we integrate dilated convolution into CRN architecture, and adopt three different types of bottleneck modules, namely long short-term memory (LSTM), gated recurrent units (GRU) and dilated convolutions. The experimental results show that the model with dilated convolution in the encoder-decoder and the model with dilated convolution at bottleneck layer have the highest PESQ and STOI scores, respectively.

1. Introduction
Speech enhancement is the restoration of clean speech from noisy signals. In other words, the purpose of speech enhancement is to separate the target speech from background interference, which may include non-speech noise and room reverberation[1]. Speech enhancement plays an important role in many real-time applications, such as robust automatic speech recognition, hearing aids[2] and mobile communication. It is necessary to ensure the quality and intelligibility of enhanced speech. The computational complexity of the algorithm should be minimized so that processed speech can be played almost immediately.

Over the past few years, deep learning-based algorithms have gained tremendous popularity and achieved state-of-the-art performance in the field of speech enhancement[3]. However, there is a problem that some excellent models either have numerous training parameters or too complex architecture, which slow down the calculation process. In order to meet the demand of real-time processing, it is necessary to simplify the model and keep its performance to a satisfactory level. Tan and others proposed a novel convolutional recurrent network (CRN) to address real-time monaural speech enhancement[4]. This model has less training parameters and better performance than LSTM model[5]. Unfortunately, the speech after processing appears to be notably distorted. The reason is that the CRN uses the magnitude spectra of speech as the training target, which leads to the neglect of phase information recovery. When signal-to-noise ratio is low, the recovery of speech signal using noisy phase and even the ground truth magnitude spectrum will lead to largely distorted speech wave form. Hao et al.[6] proposed a model based on U-Net[7] operating directly in the time domain and achieved a good performance, especially at low signal-to-noise ratio (SNR) condition. Recently, Pirhosseinloo et al.[8] proposed a dilated convolutional recurrent neural network for real-time monaural speech enhancement and this model has good generalization to untrained speakers and
unseen noises. Li et al.[9] proposed a CRN with dilated convolutional kernel to expand receptive fields.

In this paper, we focus on the enhancement of speech signals captured by fiber-optic microphones in real-time processing. Fiber-optic microphone is a kind of microphone which converts sound waves into optical signals. Some optical parameters (such as light intensity, phase, polarization state) will be changed after the sound caused vibration, and the sound signal will be restored by fiber sensor and demodulation system[10]. But after the demodulation, both stable and non-stable noise are present, accompanied by occasional interfering speech. And there are other concerning factors, such as the room reverberation, the frequency response characteristics of the optical probe. Therefore, it is difficult for the receiver to accurately restore the original speech signal. In order to obtain high quality enhanced speech, we make the CRN model run directly in the time domain.

In order to find a better performance model, we replace general convolution layers with dilated convolution layers in the encoder-decoder structure. In the absence of more trainable parameters, the receptive field increases and long temporal context is captured. Then we compared different setups of the time domain model by varying the number of layers for ED and choice of bottleneck modules which include long short-term memory (LSTM), gated recurrent units (GRU) and dilated convolutions. In addition, we also investigated the effect of increasing the number of bottleneck layers.

2. Model description

2.1. Baseline CRN
The time-frequency domain CRN proposed in [4] was adopted as our baseline model, it consists of the encoder-decoder and the bottleneck. The encoder has five convolution layers while the decoder has five transposed convolution layers. Each convolution layer in the encoder halves the frequency dimension in the time-frequency feature maps while double the channel number (number of feature maps), and each transposed convolution layer in the decoder does the opposite. It additionally incorporates skip connections to facilitate optimization, which connect each layer in the encoder to its corresponding layer in the decoder. And the bottleneck is made up of two LSTM layers.

2.2. Dilated CRN
Baseline CRN is to map the magnitude spectrogram of noisy speech to that of clean speech, where the magnitude spectrogram is treated as an image. This leads to the loss of the phase information of the original speech, thus distortion in recovered speech. In this regard, all the models we proposed are trained in the time domain.

In CRN structure, with the strided convolution and fixed kernel size in each layer of the encoder, the receptive field of the convolution kernel decreases gradually. Despite the presence of skip connections, a lot of detailed information is lost throughout the encoding process. The dilated convolution[11] can expand the receptive field without increasing training parameters. Therefore, dilated convolutions are incorporated into the encoder and decoder structure of our proposed dilated CRN (DCRN) instead of the regular ones. And the convolution kernel of each convolution layer is of size 7. The dilation coefficient grows exponentially in the encoder and decreases exponentially in the decoder. In the encoding stage, to leverage a larger context along the time dimension, we utilize max pooling to halve the length of feature tensor after each dilated convolution operation. The size of the pooling window is (1,2,1). The stride of the pooling layer and the convolution layer are (1,2,1) and 1 respectively. The decoder directly adopts one-dimensional dilated transposed convolution layers with a stride of 2. The details of DCRN architecture are given in table 1 (Here f represents the number of convolution kernels, and d represents the dilation coefficient).

2.3. CRN with different bottleneck
For exploring the model with better performance and obtain higher quality speech, we increased the complexity of the model. The number of convolution and deconvolution layers of the encoder and
decoder increased from five to seven. We added dropout to each convolution layer avoiding the problem of overfitting. The batch normalization (BN) is added to each convolution layer and the activation function uses LeakyReLU ($\text{LReLU}$)[12]. The new architecture we proposed is shown in figure 1 (a) and the detailed description is given in table 2. The size of convolution kernel is 15 and the stride is 2. At the same time, to study how different bottleneck modules affect the results, we used three different functioning blocks as the bottleneck for comparison.

Table 1. Architecture of our proposed DCRN.

| Block name   | Layer name | Hyper parameters | Output size     |
|--------------|------------|------------------|----------------|
| Reshape_1    | -          | -                | (M, 16384, 1)  |
| Encoder      | conv1d_1   | f = 16, d = 1    | (M, 8192, 16)  |
|              | conv1d_2   | f = 32, d = 2    | (M, 4096, 32)  |
|              | conv1d_3   | f = 64, d = 4    | (M, 2048, 64)  |
|              | conv1d_4   | f = 128, d = 8   | (M, 1024, 128) |
|              | conv1d_5   | f = 256, d = 16  | (M, 512, 256)  |
| Bottleneck   | LSTM_1     | 1024             | (M, 512, 1024) |
|              | LSTM_1     | 1024             | (M, 512, 1024) |
| Decoder      | deconv1d_1 | f = 128, d = 8   | (M, 1024, 128) |
|              | deconv1d_2 | f = 64, d = 4    | (M, 2048, 64)  |
|              | deconv1d_3 | f = 32, d = 2    | (M, 4096, 32)  |
|              | deconv1d_4 | f = 16, d = 1    | (M, 8192, 16)  |
|              | deconv1d_5 | f = 1            | (M, 16384, 1)  |
| Reshape_2    | -          | -                | (M, 16384)     |

Table 2. Architecture of our proposed new CRN.

| Block name   | Layer name | Hyper parameters | Output size     |
|--------------|------------|------------------|----------------|
| Reshape_1    | -          | -                | (M, 16384, 1)  |
| Encoder      | conv1d_1   | f = 16           | (M, 8192, 16)  |
|              | conv1d_2   | f = 32           | (M, 4096, 32)  |
|              | conv1d_3   | f = 64           | (M, 2048, 64)  |
|              | conv1d_4   | f = 128          | (M, 1024, 128) |
|              | conv1d_5   | f = 256          | (M, 512, 256)  |
|              | conv1d_6   | f = 512          | (M, 256, 512)  |
|              | conv1d_7   | f = 1024         | (M, 128, 1024) |
| Bottleneck   | -          | -                | (M, 128, 1024) |
| Decoder      | deconv1d_1 | f = 512          | (M, 256, 512)  |
|              | deconv1d_2 | f = 256          | (M, 512, 256)  |
|              | deconv1d_3 | f = 128          | (M, 1024, 128) |
|              | deconv1d_4 | f = 64           | (M, 2048, 64)  |
|              | deconv1d_5 | f = 32           | (M, 4096, 32)  |
|              | deconv1d_6 | f = 16           | (M, 8192, 16)  |
|              | deconv1d_7 | f = 1            | (M, 16384, 1)  |
| Reshape_2    | -          | -                | (M, 16384)     |

For the sake of tracking a target speaker, long-term context is important. But it cannot be leveraged by the convolutional encoder-decoder[8]. LSTM is a special type of recurrent neural network (RNN)[13], which incorporates input gate, forget gate and output gate. It is capable of maintaining connections within long data sequence, thus good for processing long-term sequential information.
GRU[14], a variant of LSTM, combines the forget gate and the input gate into a single update gate, as well as a mix of cell state and hidden state. GRU has relatively fewer parameters and it is easier to converge. In addition, dilated convolutions can also systematically aggregate multi-scale contextual information. The advantage of dilated convolutions is that the receptive field can be expanded without losing resolution, so that each convolution output contains a large range of information. The concrete structures of bottlenecks are explained as follows:

- **LSTM**: Firstly, we inserted two stacked LSTM layers between the encoder-decoder for modeling temporal dynamics of speech. The output dimension is 1024 for every LSTM layer. Next, we added another layer of LSTM to observe if there was a significant improvement.
- **GRU**: Similarly, we stacked two GRU layers as the bottleneck in the first. Each GRU layer has 1024 channels. Then we tried three GRU layers.
- **Dilated convolutions**: Another type of the bottleneck consists of two dilated convolution layers. The structure of the module is shown in figure 1 (b).

### Figure 1. (a) Architecture of our proposed new CRN, (b) structure of the bottleneck consisting of dilated convolutions.

#### 3. Experiments

**3.1. Dataset**

In this study, our dataset was built by the AISHELL-1 corpus[15] to evaluate our proposed models. We randomly selected 8000 utterances from AISHELL-1 corpus and divided them into three parts. There were 6480 utterances in training set, 720 utterances in validation set and 800 utterances in test set. All utterances were played in the office through high quality speaker, and a fiber-optic microphone was used to capture speech signals. Finally, after some cut and alignment we got about 15 hours of paired clean and noisy utterances in total.

**3.2. Experimental setup**

In our experiments, all the speech signals were sampled at 16 kHz. In baseline CRN, the spectrum was extracted by using a Hamming window. The frame used for short time Fourier transform has a window length of 20 ms with 50% overlap between two consecutive windows. Finally, the amplitude of the spectrum was normalized. During training, we modified the size of the convolution kernel to (7, 7) and the channels of LSTM to 1028. The activation function was changed to the LeakyReLU. The initial learning rate was 0.0002 and the batch size was 16. Other hyper parameters were consistent with the original paper.

For our proposed model, the input and output were raw speech waveforms. All speech utterances were padded with zeros to the same length. And the input utterances were divided into segments slightly longer than 1 second. Each segment contains 16,384 samples. We set the batch size to 8 and the learning rate to 0.0005.
All the models were trained with the Adam optimizer[16]. The mean squared error (MSE) served as the loss function. The best models were selected when the loss reached the minimum on the validation set.

3.3. Results
In this study, PESQ[17] and STOI[18] were used as the evaluation metrics to evaluate the performance of different models. We calculated both metrics on each utterance and averaged over test set. Table 3 presents PESQ and STOI scores of captured and enhanced signals. Captured represents the unprocessed speech signals. Baseline denotes the baseline CRN and WCRN represents baseline CRN working in the time domain. Then DCRN is our proposed dilated CRN. Next, CRNL2 denotes our proposed new CRN with two LSTM layers as the bottleneck and CRNL3 has three LSTM layers in the bottleneck. In the same way, CRNG2 and CRNG3 denotes two and three GRU layers at the bottleneck. Finally, CRND2 represents the bottleneck consisting of two dilated convolution layers.

Table 3. Average PESQ and STOI scores of different models on test set.

| Model   | PESQ    | STOI    |
|---------|---------|---------|
| Captured| 1.8745  | 0.3698  |
| Baseline| 2.3127  | 0.4292  |
| WCRN    | 2.7328  | 0.4615  |
| DCRN    | 2.9015  | 0.4785  |
| CRNL2   | 3.0787  | 0.4677  |
| CRNL3   | 3.0796  | 0.4693  |
| CRNG2   | 3.0743  | 0.4665  |
| CRNG3   | 3.0502  | 0.4667  |
| CRND2   | 3.0814  | 0.4694  |

Compared with Baseline, the performance of all new models has been improved. And WCRN gains a significant improvement of 0.42 on PESQ and 0.03 on STOI. It can be seen that the improvement is very obvious after the shift from CRN to the time domain. This indicates that WCRN network can alleviate the effect of phase information loss in baseline CRN. By comparing WCRN and DCRN, DCRN shows a slight improvement in both PESQ and STOI so that the use of dilated convolutions within the encoder-decoder can improve the performance of the model.

The CRNL2, with more encoder-decoder layers than WCRN, has an obvious improvement of 0.35 on PESQ and performs better on STOI. It shows that adding convolution layers is useful. However, when we increase the number of LSTM layers, there is not much improvement in performance. Therefore, it is not feasible to optimize the model by increasing the layers of LSTM. When the bottleneck layer is switched to the GRU, we found that the scores of PESQ and STOI are not significantly distinct. Thus there is almost no performance difference between GRU and LSTM, but GRU has fewer training parameters. The training parameters of each model are shown in figure 2.

As we can see, DCRN has the best score on STOI and CRND2 performed the best on PESQ metric. Nevertheless, CRND2 has the most training parameters and the parameter of CRND2 was nearly four times that of DCRN. Moreover, CRND2 has only a negligible improvement over CRNG2 on both PESQ and STOI. In summary, it can be concluded that, for the application of real-time processing, DCRN is the best choice if computing hardware resources are limited. If you have a higher demand for the quality of enhanced speech, CRND2 is the best choice.

In order to visualize the processing effect of each model, we chose an utterance from the test set as an example. The enhanced spectrograms of the utterance are illustrated in figure 3 for different models. The captured signal is lossy in term of high frequency component. And there is 4 kHz single-frequency interference. The baseline model removes most of the noise, but loses most of the high frequency signal. And the spectrogram of Baseline looks like being truncated at about 4000Hz. The
same problem also exists when time domain model is applied. But the spectrum of WCRN is significantly cleaner than that of Baseline. This truncation problem is solved when the layers of the encoder-decoder are increased. In addition, models with different types of bottleneck do not show significant differences.

Figure 2. The number of trainable parameters in different models.
4. Conclusions
In this paper, we focus on the research of deep learning speech enhancement in fiber-optic microphones. Different from most other speech enhancement studies, the captured speech signal is contaminated by non-additive noise instead of additive noise. We propose several improved CRN architecture, all outperforming the baseline CRN on both PESQ and STOI. We proved that baseline CRN working in the time domain can improve the performance of the model. Besides, we found that the adoption of dilated convolution layers in the encoder-decoder has the highest score on PESQ due to the larger receptive field. Finally, by comparing the experimental results of the CRN with different bottleneck modules, it is found that the performance difference of LSTM, GRU and dilated convolutions can be neglected. Furthermore, with the increase of GRU and LSTM layers, the model performance does not improve significantly.

Acknowledgments
The authors would like to thank Hefei Comprehensive National Science Center and National Key R & D Plan of China (Grant No.2016YFC0301902).

References
[1] Wang, D., Chen, J. (2018) Supervised Speech Separation Based on Deep Learning: An Overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 10:1702-1726.
[2] Bogaert, T. V., Doclo, S., Wouters, J., Moonen, M. (2009) Speech enhancement with multichannel Wiener filter techniques in multimicrophone binaural hearing aids. Journal of the Acoustical Society of America, 125(1): 360-71.
[3] Liu, C., Ai Y., Ling, Z., (2021) Phase Spectrum Recovery for Enhancing Low-Quality Speech Captured by Laser Microphones. In: 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP). Hong Kong. pp. 1-5.
[4] Tan K., Wang, D. L., (2018) A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement. In: INTERSPEECH. Hyderabad. pp. 3229-3233.
[5] Chen, J., Wang, D. L., (2017) Longshort-term memory for speaker generalization in supervised speech separation. Journal of the Acoustical Society of America, 141.6:4705.

[6] Hao, X., Su, X., Wang, Z., Zhang, H., Batushiren, A. (2019) UNetGAN: A robust speech enhancement approach in time domain for extremely low signal-to-noise ratio condition. In: INTERSPEECH. Graz. pp. 1786–1790.

[7] Ronneberger, O., Fischer, P., Brox, T. (2015) U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer. pp. 234–241.

[8] Pirhosseinloo, S., Brumberg, J. S. (2019) Dilated convolutional recurrent neural network for monaural speech enhancement. In: 2019 53rd Asilomar Conference on Signals, Systems, and Computers. Pacific Grove. pp. 158-162.

[9] Li, Y., Liu, M., Drossos, K., Virtanen, T. (2020) Sound Event Detection Via Dilated Convolutional Recurrent Neural Networks. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona. pp. 286-290.

[10] Wu, Y., Gan, J., Li, Q., Zhang, Z., Heng, X., Yang, Z. (2015) Distributed Fiber Voice Sensor Based on Phase-Sensitive Optical Time-Domain Reflectometry. IEEE Photonics Journal, vol. 7, no. 6:1-10.

[11] Yu, F., Koltun, V., (2016) Multi-Scale Context Aggregation by Dilated Convolutions. In: Proceedings of the International Conference on Learning Representations (ICLR). San Juan.

[12] Zhang, X., Zou, Y., Shi, W. (2017) Dilated convolution neural network with LeakyReLU for environmental sound classification. In: 2017 22nd International Conference on Digital Signal Processing (DSP). London. pp. 1-5.

[13] Hochreiter, S., Schmidhuber, J. (1997) Long Short-Term Memory. Neural Computation, vol. 9, no. 8: pp. 1735-1780.

[14] Cho, K., Merrienboer, B. V., Gulcehre, C., Ba Hdanau, D., Bougares, F., Schwenk, H., et al. (2014) Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Computer Science.

[15] Bu, H., Du, J., Na, X., Wu, B., Zheng, H. (2017) AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline. In: 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). Seoul. pp. 1-5.

[16] Kingma, D. P., Ba, J. (2014) Adam: A method for stochastic optimization. Computer Science.

[17] Rix, A. W., Beerends, J. G., Hollier M. P., Hekstra, A. P. (2001) 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). Salt Lake City. pp. 749-752 vol.2.

[18] Taal, C. H., Hendriks, R. C., Heudsens, R., Jensen, J. (2011) 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.