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

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

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

*whhit173@hotmail.com

Abstract

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

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

1. Introduction

In most applications, speech enhancement is concerned with improving the quality and intelligibility of speech that has been degraded by additive noise [1]. Speech enhancement algorithms not only improve some perceptual aspect of speech but also can be used as a preprocessing step for other speech-related tasks, such as automatic speech recognition [2], speech coding [3] and hearing aids [4]. Traditional speech enhancement methods have been studied for decades, and most of these methods are based on the additional nature of the speech signal and significantly outperforms the DNN with uniform assumptions [1].

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

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

http://dx.doi.org/10.21437/Interspeech.2019-2472
in one layer, but it brings a lot of parameters and increases the difficulty of network optimization. The fully-connected layer also requires a fixed-size input/output and loses positional correspondence. These shortcomings limit the use of the fully-connected layer in many cases. To address the above issues of convolutional operations, a non-local module [27] has been proposed recently and achieved impressive performance after the integration with CNNs in many computer vision tasks [27–30]. In this work, we propose a convolutional neural network with the non-local module for speech enhancement. The non-local operations calculate the mutual similarity between time-frequency units in each frame, which is helpful for capturing the global information in the frequency domain with a slightly increased computing complexity. The experimental results show that the proposed scheme improves the computational efficiency significantly and produces satisfactory enhancement performance comparing the DNN, long short-term memory (LSTM) and C-RNN baselines.

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

2. Non-local convolutional neural network for speech enhancement

2.1. Non-Local block for speech enhancement

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

\[
y_i = \frac{1}{C(x)} \sum_{x_j} f(x_i, x_j) g(x_j), \tag{1}
\]

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

\[
f(x_i, x_j) = e^{\beta(x_i)^T \phi(x_j)}, \tag{2}
\]

where \(\beta(x_i)\) and \(\phi(x_j)\) are two embedding schemes using a convolution with kernel size 1. In this case the normalization factor in Eq. (1) is set as \(C(x) = \sum f(x_i, x_j)\) and \(f(x_i, x_j) / C(x)\) becomes the softmax operation along the dimension \(j\).

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

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

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

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

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

2.2. Network configurations

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

Figure 2: Overview of the proposed architecture for speech enhancement

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

<table>
<thead>
<tr>
<th>Layers</th>
<th>Input size</th>
<th>Parameters</th>
<th>Output size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input-layer</td>
<td>(7, 129)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Extension Block</td>
<td>(7, 129)</td>
<td>(32, 1, 1)</td>
<td>(256, 32)</td>
</tr>
<tr>
<td>Convolution Layer</td>
<td>(256, 32)</td>
<td>(32, 3, 1), 4</td>
<td>(256, 32)</td>
</tr>
<tr>
<td>Convolution Layer</td>
<td>(256, 32)</td>
<td>(32, 1, 1)</td>
<td>(256, 32)</td>
</tr>
<tr>
<td>Non-Local Block</td>
<td>(256, 32)</td>
<td>(32, 1, 1), 2</td>
<td>(256, 32)</td>
</tr>
<tr>
<td>Flatten</td>
<td>(256, 2)</td>
<td>(1, 1)</td>
<td>(512,)</td>
</tr>
<tr>
<td>Fully-connected Layer</td>
<td>(512,)</td>
<td>(129,)</td>
<td>(129,)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Metrics</th>
<th>PESQ</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Noisy +1 NL</td>
<td>Noisy +2 NL</td>
</tr>
<tr>
<td>babble</td>
<td>2.0919</td>
<td>2.4979</td>
</tr>
<tr>
<td>factory1</td>
<td>1.9882</td>
<td>2.5712</td>
</tr>
<tr>
<td>railway</td>
<td>2.0519</td>
<td>2.6237</td>
</tr>
<tr>
<td>white</td>
<td>1.7270</td>
<td>2.5753</td>
</tr>
<tr>
<td>Avg.</td>
<td>2.1982</td>
<td>2.5670</td>
</tr>
</tbody>
</table>

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

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

3. Experiments

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

1. DNN [11]. DNN contains 3 hidden layers of size 1024 and sigmoid activation functions are applied to all fully-connected layer except the output layer.
2. LSTM [20]. LSTM contains 2 hidden layers, both of which has 1024 hidden units. And the outputs of LSTMs are fed into one fully-connected layer.
3. C-RNN [25]. The convolutional component uses a 2-D convolution with 64 filters, kernel size (T, 16) and time-frequency stride (1, 8). Two layers of bidirectional LSTMs follow the convolutional component, each of which has 512 hidden unit.
The input context window of each method is set as 11 \((T = 11)\) and it spans from past 5 frames to future 5 frames. We use Adam optimizer [36] for the training and the initial learning rate is set to 0.001 with \(\beta_1 = 0.9\), \(\beta_2 = 0.999\) and \(\epsilon = 1.0e^{-8}\). The mini-batch size is set to \(N = 128\) and the loss function is mean square error (MSE). Total number of epoch is 100. When the validation loss doesn’t decrease for more than 5 epochs, the training will be terminated in advance to prevent overfitting. We select the model which produces the best performance on the validation set for experiments. The objective speech quality and intelligibility is evaluated via perceptual evaluation of speech quality (PESQ) [37] and Short-Time Objective Intelligibility (STOI) [38] scores, respectively.

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

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

![Table 3: PESQ and STOI values of the different methods on all noises types.](Image)

<table>
<thead>
<tr>
<th>Metrics</th>
<th>PESQ</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Noisy</td>
<td>DNN</td>
</tr>
<tr>
<td>SNR 15</td>
<td>2.7007</td>
<td>2.8067</td>
</tr>
<tr>
<td>SNR 10</td>
<td>2.3771</td>
<td>2.6743</td>
</tr>
<tr>
<td>SNR 5</td>
<td>2.0415</td>
<td>2.4715</td>
</tr>
<tr>
<td>SNR 3</td>
<td>1.9110</td>
<td>2.3601</td>
</tr>
<tr>
<td>SNR 0</td>
<td>1.7254</td>
<td>2.1744</td>
</tr>
<tr>
<td>SNR -3</td>
<td>1.5503</td>
<td>1.9496</td>
</tr>
<tr>
<td>SNR -5</td>
<td>1.4532</td>
<td>1.7904</td>
</tr>
<tr>
<td>Avg.</td>
<td>1.9648</td>
<td>2.3181</td>
</tr>
</tbody>
</table>

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

![Table 4: The number of parameters between the baselines and proposed model.](Image)

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of parameters (Million)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNN</td>
<td>3.68</td>
</tr>
<tr>
<td>LSTM</td>
<td>13.25</td>
</tr>
<tr>
<td>C-RNN</td>
<td>12.47</td>
</tr>
<tr>
<td>NL</td>
<td>0.13</td>
</tr>
</tbody>
</table>

4. Conclusions

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


