Monoaural Speech Enhancement Using a Nested U-Net with Two-Level Skip Connections

Seorim Hwang\textsuperscript{1}, Sung Wook Park\textsuperscript{2}, Youngcheol Park\textsuperscript{1}

\textsuperscript{1}Intelligent Signal Processing Lab, Yonsei University, Wonju, Korea.
\textsuperscript{2}Gangneung-Wonju National University, Gangneung, Korea.

Abstract

Capturing the contextual information in multi-scale is known to be beneficial for improving the performance of DNN-based speech enhancement (SE) models. This paper proposes a new SE model, called NUNet-TLS, having two-level skip connections between the residual U-Blocks nested in each layer of a large U-Net structure. The proposed model also has a causal time-frequency attention (CTFA) at the output of the residual U-Block to boost dynamic representation of the speech content in multi-scale. Even having the two-level skip connections, the proposed model slightly increases the network parameters, but the performance improvement is significant. Experimental results show that the proposed NUNet-TLS has superior performance in various objective evaluation metrics to other state-of-the-art models. The code of our model is available at https://github.com/seorim0/NUNet-TLS

Index Terms: speech enhancement, multi-scale feature, attention mechanism

1. Introduction

Speech enhancement (SE) is to suppress the background noise and, as a result, improves the quality and intelligibility of noisy speech signals \cite{1}. The noise reduction process can be used as a significant factor for other speech-related applications such as speech recognition \cite{2} and speech coding \cite{3}. The conventional single-channel speech enhancement methods attempt to improve the sound quality based on input statistics \cite{4–6}. Naturally, the best performance can be obtained when the input statistics are known, or the assumptions used in the method are matched well with the real-world conditions \cite{7}. To overcome this limitation, deep neural network (DNN)-based SE utilizing a vast amount of data has been widely and successfully tried. The supervised learning-based DNN outperformed the conventional approach without any assumptions about the statistical properties of the input \cite{8–10}. For this purpose, various types of real or complex networks have been proposed \cite{11, 12}.

Previous studies \cite{9, 10, 13, 14} showed that proper use of contextual information in speech could improve the performance of the SE model. Since speech signals have global and local contexts, capturing and utilizing both can improve SE performance. Based on this, various methods of capturing contextual information in multiple scales have been proposed by using different kernel sizes \cite{10}, adjusting the dilation size of kernels \cite{13}, or using a separate network \cite{14}. However, most of the previous approaches required a significant increase in computational complexity not to compromise the accuracy of the contextual information.

U\textsuperscript{2}-Net \cite{8} was proposed to efficiently extract multi-scale features with a low computational cost for salient object detection. It has a large U-Net structure with U-Net-shaped residual blocks in each layer, allowing us to capture contextual information at different scales. Inspired by U\textsuperscript{2}-Net, a nested U-Net with self-attention and dense connectivity (SADNUNet) was recently proposed \cite{9} for the monaural SE task. It takes a time-domain frame as input and captures multi-scale temporal features using residual U-blocks nested on a large U-Net. Also, at the bottleneck, it uses a DenseNet \cite{15} and self-attention to further capture the global contextual feature with high accuracy. Although U\textsuperscript{2}-Net and SADNUNet showed excellent performance in each application, they did not fully utilize the capabilities of nested U-Net structures. Since the skip connection in the U-Net structure is known to play an essential role in achieving high performance by compensating the detailed context, there are still possibilities for improving the performance by developing different skip connections. In addition, it is also possible to further boost the SE performance by employing a more sophisticated attention strategy.

This paper proposes a nested U-Net with two-level skip connections (NUNet-TLS), where layer-by-layer long skip connections are established in two levels to pass the features from the encoder part to the decoder part. In addition to top-level skip connections, as found in the original U-Net, the proposed NUNet-TLS uses inner-level long skip connections between the residual U-blocks nested on each layer in the encoding and decoding paths. In such a way, the proposed NUNet-TLS can recover local spectra information lost during downsampling with higher accuracy than U\textsuperscript{2}-Net and SADNUNet. We also employ a causal T-F attention (CTFA) within the residual U-block to effectively boost the dynamic representation of the contextual information in the time-frequency domain.

2. Backgrounds

Recognizing the importance of contextual information, many DNN-based SE models attempt to take full advantage of contextual information, comprising the global and local contexts of the input speech signal. These attempts can be divided into two main streams: 1) expanding receptive fields with different size kernels \cite{10, 13, 14} and 2) using additional sub-networks \cite{14}. U\textsuperscript{2}-Net \cite{8} possesses both of those two aspects. It has a U-Net-like structure, and the layers in the U-Net are made up of different sizes of U-Net-shaped residual blocks. Repeated downsampling and upsampling of the input data are equivalent to adjusting the receptive field. And since it is a nested two-level structure, it can be considered a method of using a sub-network.

SADNUNet \cite{9}, on the other hand, inherits the U\textsuperscript{2}-Net architecture and successfully improves the SE performance after a few modifications. It accepted the time-domain speech frames and employed sub-pixel convolutions \cite{16} for the up-sampling in the decoder path. It also replaced the bottleneck’s dilated blocks of U\textsuperscript{2}-Net with dilated dense (DDense) blocks and added...
a self-attention to the input layer and top-level bottleneck. Since the dense block at the bottleneck enhances gradient flow, it can improve the model performance.

Skip connection is commonly used for providing local context in the encoder-decoder structure [17–19]. In the U-Net, the decoder upsamples the encoded data and reconstructs or enhances the input data. U²-Net and SADNUNet comprises residual U-blocks for the additional multi-scale capturing of the contextual features, and both use standard skip connections in the top-level structure. Although the top-level skip connections can compensate for the lost features during the encoder’s down-sampling, it still does not sufficiently compensate for the local context of the encoder-decoder process. The residual U-block in the decoder part depends only on the input provided through the top-level from the encoding part.

The attention mechanisms have been widely studied in various fields to allow DNNs to learn what is essential to their learning task [20]. Early attention mechanisms for SE mainly focused on effectively modeling long-distance dependencies. However, since it is important to interpret contextual information of speech between different time-frequency (T-F) units, recently proposed attentions [21, 22] utilize the two-dimensional information in the time and frequency axis.

3. Proposed SE Model

This section presents the overall structure of our nested U-Net with two-level skip connections (NUNet-TLS) and the residual U-block comprising a causal time-frequency attention.

3.1. Model Architecture

The overall architecture of the proposed NUNet-TLS is presented in Fig. 1. It works on the time-frequency domain and consists of four components: one input layer, six symmetric encoder-decoder stages, one bottleneck block, and one output layer. The encoder and decoder stages comprise a modified residual U-block to extract multi-scale features together with a causal time-frequency attention in both encoder and decoder part. In the encoder stage, a down-sampling layer follows the modified residual U-block, and in the decoder stage, an up-sampling layer precedes the modified residual U-block, respectively. And the bottleneck consists of a dilated dense (DDense) block, similar to SADNUNet [9].

The network first converts the noisy input $y$ to T-F domain magnitudes $|Y|$ using the convolutional short-time Fourier transform (ConvSTFT) module. Then it subsequently goes through Multi-Scale Feature Extraction (MSFE) blocks in the encoder, a DDense block, and MSFE blocks in the decoder. In Fig. 1, MSFE blocks in the encoder and decoder parts are distinguished as MSFEe and MSFEd, respectively, because they have different dimensions due to the inner-level skip connections. Numbers inside the MSFE blocks in Fig. 1 indicate the number of layers of the residual U-block, i.e., $i$ of MSFEe-i or MSFEd-i. The output layer obtains the enhanced magnitude spectra $|\hat{X}|$ which is combined with the phase of the noisy input $\theta_y$ to obtain the enhanced output $\hat{X}$. Finally, the convolutional ISTFT (ConvISTFT) recovers a frame of time-domain output $\hat{x}$.

The MSFE block has a similar structure to the residual U-block in [8] and [9], but there are differences, as will be explained in the following subsections.

\begin{equation}
\hat{x} = \mathcal{U}(f(\hat{X}_n) \odot A_{tf} + f(\hat{X}_{n-1}))
\end{equation}

where, $f$ denotes convolution operation, $\mathcal{U}$ denotes the multi-scale feature extraction operation, $A_{tf}$ is the causal TFA, and $\odot$ denotes the element-wise multiplication operation. For convenience, $\hat{X}_n^e$ and $\hat{X}_n^d$ are regarded as $\hat{X}_n$.
Figure 2: Schematics of the two-level skip connections between Multi-Scale Feature Extraction (MSFE) blocks in the encoder path (e-i) and the decoder path (d-i). Parameters in the layer block indicate the input channel size ($C$ with subscripts in, mid, and out), the kernel size (stride along time, stride along frequency), and the output channel size.

where $n$ and averaged frequency data are obtained as convolution and nonlinear activation layers. The averaged time and frequency data are simultaneously utilized. The TA and FA are obtained by modifying the TFA in [22] where a time-attention (TA) working in the time-domain mean-squared-error (MSE) is commonly used as a loss term to train DNN models for SE. However, MSE alone often shows a limited performance [24]. As a supplement to the time-domain MSE, an additional frequency-domain loss function is widely used [9, 11, 25]. This paper also uses a joint time-frequency loss function, as given by

$$L = \lambda_t L_t + \lambda_f L_f,$$

where $L_t$ and $L_f$ are the frequency-domain magnitudes of $x_t$ and $x_f$.

3.4. Loss function

The key to TFA is to properly apply the energy distribution of speech in the time and frequency dimensions to the attention map. The CTFA used in this paper is obtained by modifying the stacked layers. In addition, the DDense block comprises no self-attention in both large U-Net and MSFE blocks. These differences resulted in an improvement of the SE performance, as shown in the Experiments.

3.3. Causal Time-Frequency Attention (CFTA)

The proposed MSFE block in Fig. 2 by itself is different from the previous residual U-Net blocks in [8] and [9] in several aspects: first, it is working in the time-frequency domain. Second, it has a causal time-frequency attention on the top of the stacked layers. In addition, the DDense block comprises no self-attention in both large U-Net and MSFE blocks. These differences resulted in an improvement of the SE performance, as shown in the Experiments.

where $l$ and $k$ are the time-stamp and frequency bin indices, respectively, with their maximum ranges of $L$ and $F$. One major shortcoming of this approach is that causality is not guaranteed since it operates per block of frames. To suffice the causality necessary for real-time implementation, we modify the equation for $A_l(k)$ so that it includes only the causal frames:

$$A_l(k, l) = \frac{1}{P} \sum_{p=1}^{P} G_{n,i}(l - p, k),$$

where $P$ is the number of look-back frames, and in this paper, it was experimentally set to $P = 32$. Zero-padding is used when $p > l$. Finally, a 2-D TFA map $A_{l,f}$ is obtained through tensor multiplication of the 1-D TA using $A_l(l)$ and 2-D FA using $A_f(k, l)$.

4. Experiments

4.1. Datasets and test setups

For experiments, the TIMIT corpus [26] was used as a clean speech dataset. We randomly selected 3,696 utterances from 630 female and male speakers for training and 462 for testing.
These utterances for training were mixed with a set of noises at signal-to-noise ratio (SNR) conditions of 0 dB to 15 dB at 1 dB intervals, respectively. And the utterances for test were mixed at 0, 5, 10, and 15 dB SNR conditions. Ten types of noises were chosen for training from CHIME-2 [27], CHIME-3 [28], and NOISEX-92 [29] databases, and the test noises were selected from the ETSI [30] database. In total, we used 59,136 noisy speech clips for training and 1,848 noisy speech clips for test. The sampling rate of all signals and noises was 16 kHz. Window length, hop size, and FFT length were 25ms, 6.25ms, and 512 samples.

We compared the proposed model (NUNet-TLS) with other recent SE models, including DCCRN with complex convolutional block attention module (DCCRN+C) [25], FullSubNet [14], and SADNUNet [9]. Total parameters of DCCRN+C, FullSubNet, and SADNUNet were 3.77M, 5.64M, and 2.63M, respectively, and the proposed model has 2.83M parameters. Thus, the proposed model doesn’t significantly increase the number of parameters. We used Adam optimizer, and the learning rate was 0.001 and 0.0001. To evaluate and compare the performance of the tested methods, four objective parameters were measured, including perceptual evaluation of speech quality (PESQ) [31], CSIG, CBAK, and COVL [32].

4.2. Results and discussions

First, an ablation test was performed to evaluate the effectiveness of the various components comprised in the proposed NUNet-TLS, and the results are given in Table 1. The baseline in the table is the model shown in Fig. 1, but without the tested components such as two-level skip connections (TLS) and causal TFA (CTFA). PESQs were obtained by averaging over all SNR cases.

<table>
<thead>
<tr>
<th>Model</th>
<th>Params.</th>
<th>PESQ</th>
<th>CSIG</th>
<th>CBAK</th>
<th>COVL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy</td>
<td>1.63</td>
<td>3.52</td>
<td>2.21</td>
<td>2.89</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>2.58M</td>
<td>3.21</td>
<td>4.72</td>
<td>3.72</td>
<td>4.20</td>
</tr>
<tr>
<td>+TLS</td>
<td>+0.20M</td>
<td>3.33</td>
<td>4.77</td>
<td>3.82</td>
<td>4.29</td>
</tr>
<tr>
<td>+CTFA</td>
<td>+0.05M</td>
<td>3.29</td>
<td>4.74</td>
<td>3.79</td>
<td>4.25</td>
</tr>
<tr>
<td>NUNet-TLS</td>
<td>+0.25M</td>
<td>3.36</td>
<td>4.79</td>
<td>3.89</td>
<td>4.32</td>
</tr>
</tbody>
</table>

The PESQ improvements by employing TLS and CTFAs are 0.12 and 0.08, respectively. And the total improvement by using both TLS and CTFAs jointly is almost 0.15. Thus, it is confirmed that the two-level skip connection helps the network capture the input speech’s contextual information. The CTFAs embedded in the MSFE block also effectively boosts the dynamic representation of the speech context.

Next, we compared the performance of the proposed NUNet-TLS with other previous models. Table 2 shows the results. The proposed NUNet-TLS performed significantly better than the other compared models in all objective parameters and SNR conditions. First of all, the proposed model’s PESQ is as high as over 3.0 except for 0dB SNR. The improvement of PESQ over SADNUNet is about 0.25 on average. Compared with DCCRN+C, the PESQ improvement is more significant, about 0.33. The improvements of CSIG, CBAK, and COVL are also significant: 0.14, 0.20, and 0.20, respectively, on average, compared to SADNUNet. It is noteworthy that even our baseline model itself performs better than SADNUNet. PESQ improvement is almost 0.10 as shown in Tables 1 (Baseline model’s average PESQ 3.21) and 2 (SADNUNet’s average PESQ 3.11), which is mainly because operating in the T-F domain allows better use of contextual information of the speech. But the baseline model has slightly fewer parameters than SADNUNet by 0.05M since self-attention is not used in the DDense blocks of the large U-Net and the number of layers inside the residual U-blocks in the last stage is reduced by 3.

The superior performance of the proposed model can also be confirmed from the spectrograms shown in Fig. 3. Unlike the other models that can’t fully reconstruct the harmonic structure, as in the box of the plot, the proposed model (Fig. 3 (e)) reconstructs the detailed harmonic patterns, which results in lower speech distortion, i.e., higher objective parameter scores.

5. Conclusions

We proposed a new speech enhancement model, called NUNet-TLS, having skip connections between the residual U-Blocks nested in each layer of large U-Net structure. Together with the causal time-frequency attention (CTFA), the new skip connections enable the SE model to better capture the contextual features in multi-scale. Through experiments, we could confirm that the proposed model significantly outperforms the other models in all objective parameters and in spectrograms as well.

Table 2: Measurement results of the objective parameters: PESQ, CSIG, CBAK, and COVL.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Model</th>
<th>0dB</th>
<th>5dB</th>
<th>10dB</th>
<th>15dB</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PESQ</td>
<td>Noisy</td>
<td>1.20</td>
<td>1.41</td>
<td>1.73</td>
<td>2.18</td>
<td>1.63</td>
</tr>
<tr>
<td></td>
<td>DCCRN+C [25]</td>
<td>2.38</td>
<td>2.88</td>
<td>3.28</td>
<td>3.59</td>
<td>3.03</td>
</tr>
<tr>
<td></td>
<td>FullSubNet [14]</td>
<td>2.28</td>
<td>2.71</td>
<td>3.16</td>
<td>3.49</td>
<td>2.91</td>
</tr>
<tr>
<td></td>
<td>SADNUNet [9]</td>
<td>2.54</td>
<td>2.97</td>
<td>3.31</td>
<td>3.60</td>
<td>3.11</td>
</tr>
<tr>
<td></td>
<td>NUNet-TLS</td>
<td>3.79</td>
<td>3.32</td>
<td>3.58</td>
<td>3.88</td>
<td>3.36</td>
</tr>
<tr>
<td>CSIG</td>
<td>Noisy</td>
<td>3.92</td>
<td>3.33</td>
<td>3.70</td>
<td>4.14</td>
<td>3.52</td>
</tr>
<tr>
<td></td>
<td>DCCRN+C [25]</td>
<td>4.00</td>
<td>4.39</td>
<td>4.62</td>
<td>4.75</td>
<td>4.45</td>
</tr>
<tr>
<td></td>
<td>FullSubNet [14]</td>
<td>4.00</td>
<td>4.43</td>
<td>4.68</td>
<td>4.90</td>
<td>4.50</td>
</tr>
<tr>
<td></td>
<td>SADNUNet [9]</td>
<td>4.27</td>
<td>4.60</td>
<td>4.79</td>
<td>4.93</td>
<td>4.65</td>
</tr>
<tr>
<td></td>
<td>NUNet-TLS</td>
<td>4.44</td>
<td>4.79</td>
<td>4.93</td>
<td>4.98</td>
<td>4.79</td>
</tr>
<tr>
<td>CBAK</td>
<td>Noisy</td>
<td>1.55</td>
<td>1.95</td>
<td>2.40</td>
<td>2.94</td>
<td>2.21</td>
</tr>
<tr>
<td></td>
<td>FullSubNet [14]</td>
<td>2.91</td>
<td>3.35</td>
<td>3.75</td>
<td>4.13</td>
<td>3.56</td>
</tr>
<tr>
<td></td>
<td>SADNUNet [9]</td>
<td>3.17</td>
<td>3.55</td>
<td>3.87</td>
<td>4.16</td>
<td>3.69</td>
</tr>
<tr>
<td></td>
<td>NUNet-TLS</td>
<td>3.33</td>
<td>3.72</td>
<td>4.07</td>
<td>4.35</td>
<td>3.89</td>
</tr>
<tr>
<td>COVL</td>
<td>Noisy</td>
<td>2.37</td>
<td>2.70</td>
<td>3.04</td>
<td>3.46</td>
<td>2.89</td>
</tr>
<tr>
<td></td>
<td>DCCRN+C [25]</td>
<td>3.46</td>
<td>3.88</td>
<td>4.16</td>
<td>4.37</td>
<td>3.97</td>
</tr>
<tr>
<td></td>
<td>FullSubNet [14]</td>
<td>3.44</td>
<td>3.84</td>
<td>4.14</td>
<td>4.43</td>
<td>3.96</td>
</tr>
<tr>
<td></td>
<td>NUNet-TLS</td>
<td>3.85</td>
<td>4.24</td>
<td>4.49</td>
<td>4.71</td>
<td>4.32</td>
</tr>
</tbody>
</table>

Figure 3: Comparison of spectrograms: (a) clean speech, and enhanced speeches using (b) DCCRN+C, (c) FullSubNet, (d) SADNUNet, and (e) NUNet-TLS.
6. References


