Speech Denoising with Residual Attention U-Net

Jun Takahashi^{a,*}, Sayuri Kohmura^a and Taro Togawa^a

^aFUJITSU LABORATORIES LTD., 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa, JAPAN

*Corresponding Author: takahashi_jun@fujitsu.com

Abstract

We applied end-to-end speech denoising to remove background noises from noisy, monaural speech signals by directly processing a raw waveform. Recent approaches have demonstrated effective results using various deep neural network (DNN) architectures. We propose the residual attention U-Net, which connects the same layer of multiple stacked residual channel attention encoder/decoder models for speech denoising. We evaluated the proposed method using an unseen test with single-channel speech denoising. Both objective and subjective evaluations indicated that our proposed method is preferred to other speech denoising methods.

Keywords: speech denoising, speech enhancement, deep learning, attention network, residual network, U-Net.

1. Introduction

In the development of audio signal processing technologies, various information such as text (speech recognition), speaker (speaker verification), and emotion (emotion recognition) can be extracted from speech signals. The performance of these technologies is degraded when the SNR is less than 20 dB^(1,2). For high-accuracy information extraction, audio devices such as headsets, which can input with high SNR, has been equipped. The equipped devices limit the use case. Therefore, speech denoising⁽³⁾ (or speech enhancement) is required during preprocessing. Our goal is to develop monaural (single-channel) speech denoising technologies to improve under the real environment, which SNR is 5-10dB, to SNR 20dB or more. If this can be realized, audio devices without equipped, such as general-purpose microphones or conference microphones, can be used for audio information extraction. And it means that the scope of application of audio information extraction can be extended.

As our research focuses on monaural speech denoising. A multi-channel recode such as microphone array⁽⁴⁾ cannot be applied. Traditionally, Spectral subtraction⁽⁵⁾ and Wiener filter⁽⁶⁾ were popular algorithm that make noise reduction filter from its distribution and characteristic of the noise. Recently, a denoising method using a deep neural network (DNN) has been proposed^(7,8).

When paying attention to features using speech denoising, typical speech denoising methods estimate background noise by using spectrogram features on the speech signals. The short-time Fourier transform (STFT) is used to convert speech signals to a spectrogram. The denoising operation has also been applied to the magnitude⁽⁹⁾ and complex⁽¹⁰⁾ components of a spectrogram. A disadvantage is that these approaches depend on many parameters (window size of STFT and overlap of audio frames) that affect the time and frequency resolution. Moreover, signal artifacts arise due to time aliasing when using the inverse STFT.

To resolve this issue, end-to-end speech denoising methods that directly operate on raw waveforms have been proposed^(11,12,13). These methods need to deal with long dimensional data because speech signals contain high-density data samples. Therefore, the layers of a network become deeper, and deeper networks are more difficult to train.

Our contributions are follows: Based on the U-Net architecture⁽¹⁴⁾, which connects the same layer of multiple stacked encoder/decoder models, we added a residual attention network for each encoder/decoder. From the unseen test with single-channel speech denoising, we confirmed that our proposed method performed better than other speech denoising methods in both objective and subjective evaluations. Finally, we confirmed the improvement of information extraction such as speech recognition by the speech denoising.

2. Related Works

As speech signals have a high density of data samples (16,000 samples per second), end-to-end speech denoising methods need to deal with long dimensional data. Therefore, to train long dimensional data at high speeds, architectures based on convolutional neural networks (CNNs) such as U-Net and dilated CNN⁽¹⁵⁾ have been proposed.

The U-Net is an expanded version of an autoencoder that consists of encoder and decoder network. The encoder and decoder are multilayered and successively reduce (in half) and double the resolution of features in each layer. The middle layer has a skip connection. Although the U-Net has been originally used for semantic segmentation, it is also used in illustration conversions⁽¹⁴⁾ and the extraction of aerial images⁽¹⁶⁾. For speech denoising, SEGAN⁽¹¹⁾ has been proposed using generative adversarial networks⁽¹⁸⁾. Wave-U-Net⁽¹⁹⁾ has been proposed for speech source separation.

Dilated CNN, an architecture based on multi-stacked dilated convolutions, is operated using a filter that is applied by skipping input values at a certain step. It is equivalent to a convolution with a larger filter with smaller parameters, and speech signals that use an exponentially increased dilated filter can be applied. Denoising WaveNet⁽¹²⁾, which was developed as a speech denoising method based on WaveNet’s architecture⁽²⁰⁾, can achieve high quality text-to-speech results. Speech denoising with Deep Feature Loss⁽¹³⁾ was a proposed optimization that uses different networks trained for acoustic environment detection and domestic audio tagging.

The U-Net is a very simple network compared to dilated CNN. In the following sections, we describe the U-Net-based approach to improve the model performance.

3. Residual Attention U-Net

3.1 Definitions

As discussed above, network optimization is difficult when the layer of the network becomes deeper. To solve this problem, We adopt residual attention network^(21,22) on a U-Net architecture to improve the performance of speech denoising.

By definition, speech signals and background noise are typically formulated as $M = S + E$, where $M \in [-1, 1]^{T \times C}$ is a noisy speech signal, $S \in [-1, 1]^{T \times C}$ is a clean speech signal, $E \in [-1, 1]^{T \times C}$ is background noise, T is the

Fig. 1. Denoising model using Wave-U-Net.

number of audio samples, and C is the number of audio channels. As we are focusing on speech denoising in monaural audio, C is set to 1. The goal of speech denoising is to estimate S (and E), given M . Hereinafter, speech signals are defined as $(T, 1)$ array dimensions, where the first and second elements represent the temporal and channel axes respectively. In our evaluation, we used a speech signal of the sampling frequency 16 kHz and set $T = 16384$. This means about one second of speech.

3.2 The base architecture

Figure 1 shows our denoising model, referred to as U-Net architectures^(14,19). This model comprises an encoder, decoder, and skip connections (left block, right block and horizontal arrows in Fig. 1). “Conv(c, k, s)” in Fig. 1 means one-dimensional convolutional operation with c filters (channels), kernel size k , and s strides. To prevent size change after convolutional operations, zero padding can be applied at the edge of the temporal axis. “Downsampling” in Fig. 1 means average pooling with size 2 along the temporal axis. “Upsampling” in Fig. 1 repeats the temporal step twice along the temporal axis. “Concatenate” concatenates two inputs along the channel axis. “Block Enc/Dec” in Fig. 1 means some convolutional operations.

Our model inputs the noisy speech signals, which are mixed clean speech signals and background noise, and have the shape of $(16384, 1)$. And it outputs denoised speech signals, which have the same shape. We also calculate the

Fig. 2. Component of encoder/decoder blocks.

background noise by subtracting the denoised speech signals from the noisy speech signals, which is used for loss function described in the next subsection.

3.3 Residual Attention block

Next, we describe the component of the encoder and decoder, which are shown in Fig. 2. Figure 2(a), 2(b), and 2(c) show the normal block, residual block, and residual attention block respectively. Normal block indicates the usual U-Net architecture with two convolutional operations with batch normalization and ReLU activation (“BatchNorm” and “ReLU” in Fig. 2). Residual block applies a residual path⁽²³⁾ to a normal block to resolve vanishing gradient. We applied one-dimensional convolutional operation with c filters and size 1 kernel as a residual path. Residual attention block applies and

Table 1. Relationship of blocks and parameter c .

Block	Parameter c
Block Enc i for $i = 1, \dots, 10$	$32 \times i$
Block Enc 11	352
Block Dec i for $i = 1, \dots, 10$	$32 \times i$

adaptively learns channel-wise features in a residual block. “CA(r)” in Fig. 2(c) indicates the channel-wise attention block with parameter r . Figure 2(d) shows the channel-wise attention block when both the input and output have the shape of (t, c) . This architecture, known as a squeeze-and-excitation network⁽²⁴⁾, is composed of an information compression step and an information expansion step. “Global Pooling” in the Fig. 2(d) outputs channel-wise global temporal average $((t, c) \rightarrow (1, c))$. Convolutional operations compress and expand the channel information $((1, c) \rightarrow (1, c/r) \rightarrow (1, c))$. The channel information is set to $[0, 1]$ by the sigmoid function (“Sigmoid” in Fig. 2(d)) and performs channel-wise multiplication. Parameter r in Fig. 2(c), 2(d) means the reduction parameters that influence the precision and computational complexity. Based on previous studies⁽²⁴⁾, we set $r = 8$.

Parameter c in Fig. 2 is the number of convolutional filters used in the block. Table 1 shows the relationship between blocks and parameters. With SEGAN, the number of convolutional filters increases multiplicatively according to the hierarchy of the block, which causes the network scale to increase. Thus, we used a value that is incrementally increased to the hierarchy of the block.

Our proposed model inputs noisy speech signals and outputs denoised speech signals and background noise. We trained the model in such a way that estimated speech signals and background noise are close to true clean speech signals and true background noise. From previous works^(12,13), we defined the loss function:

$$\frac{1}{T} \sum_{t=1}^T (|s_t - \hat{s}_t| + |e_t - \hat{e}_t|)$$

where s_t is a clean speech signal, e_t is background noise, $\hat{s}_t$ is a denoised speech signal, and $\hat{e}_t$ is estimated background noise at sampling time t .

4. Experiments

4.1 Dataset

We evaluated the effectiveness of our proposed method from both objective and subjective evaluations. We used

single-channel noisy dataset created by Valentini et al.⁽²⁵⁾. We chose this data set because it is open to the public⁽²⁶⁾ and includes many types of noise for many different speakers, which fits the purposes of this work. The training data set is generated from the speech data of 28 speakers (14 males and 14 females) from the VTCK Corpus⁽²⁷⁾ and ten types of background noise (two artificially generated noises, namely speech-shaped noise and babble noise, and eight real noise recordings from the DEMAND⁽²⁸⁾ database). Each noise is used to generate four signal-to-noise ratios (SNRs) of 0, 5, 10, and 15 dB. Therefore, the training set had 40 different noise conditions. The original data set had been sampled at 48 kHz, and we resampled it at 16 kHz and converted it to monaural sound. The complete training set comprises 11,572 files of clean/noisy data. We loaded audio files and divided them into 16,384 samples. Of the total data, 90% is used for training and 10% for validation. We create true background noise by subtracting clean data from noisy data.

Each test set contained two speakers (one male, one female) and five other noises from the DEMAND database. The noise types were “living room,” “office space,” “bus,” “cafeteria,” and “public square.” None of these were used in the training set. The noises of the test set, with 824 files of clean/noisy data, were mixed with the SNRs of 2.5, 7.5, 12.5, and 17.5 dB. For the denoising, the end of noisy data were zero-padded so that the sample lengths were equal to multiples of 16,384 samples. They were applied to the model, and fractions of denoised data were removed. Therefore, the number of samples of the denoised data was equal to that of the original.

4.2 Experimental Setup

The proposed method was implemented by TensorFlow 1.14⁽²⁹⁾. For the training, we used the previously described loss function along with the ADAM optimizer with the learning rate of 0.001. The training batch size was 16 and the number of training epochs was set to 200.

4.3 Objective Evaluation

First, we evaluated the effectiveness of speech denoising by calculating the following objective measures (the higher, the better) provided in the references^(3,9): Perceptual evaluation of speech quality (**PESQ**) using the wide-band version recommended in ITU-T P.862.2, predictor of signal distortion attending only to the speech signal (**CSIG**), predictor of the intrusiveness of background noise (**CBAK**), the predictor of overall quality (**COVL**),

Fig. 3. Validation loss comparing the block type.

Table 2. Objective evaluation results comparing the block type.

Method	PESQ	CSIG	CBAK	COVL	SSNR	SNR
Normal	2.62	3.93	3.34	3.28	9.91	18.89
Residual	2.66	3.99	3.36	3.32	9.89	18.63
Proposal	2.77	4.08	3.42	3.43	10.10	19.17

segmental signal-to-noise ratio (**SSNR**) and SNR. The unit of SSNR and SNR is decibels.

Figure 3 show the result of validation loss for each block on training. The “Normal”, “Residual” and “Proposal” labels means that the component of U-Net encoder/decoder block described in Fig. 2(a), Fig. 2(b) and Fig. 2(c), respectively. The value of validation loss of “Residual” and “Proposal” lead to smaller values compare with “Normal” because of residual path in each block. And Table 2 shows the result of objective measures for each block. The “Block type” label in the table refers to the component of U-Net encoder/decoder blocks. In these results, our proposal outperforms than other methods. We think that residual attention block with U-Net leads to more performance.

Table 3 shows the result of objective measures compared with various conventional speech denoising methods. The “Noisy” label means that no denoising method was adopted. The “Wiener” label refers to the

Table 3. Objective evaluation results comparing the performance of among the method.

Method	PESQ	CSIG	CBAK	COVL	SSNR	SNR
Noisy	1.97	3.34	2.44	2.63	1.68	8.45
Wiener	2.22	3.23	2.68	2.67	5.07	12.28
SEGAN	2.36	3.63	2.92	2.96	6.39	14.43
DFL	2.52	3.86	3.28	3.18	9.86	18.86
Proposal	2.77	4.08	3.42	3.43	10.10	19.17

results of Wiener filter⁽⁶⁾. These are baseline methods. “SEGAN” and “DFL” labels imply the results of end-to-end speech denoising method, SEGAN⁽¹¹⁾ and Deep Feature Loss⁽¹³⁾, respectively. The models of these methods trained the same dataset described in 4.1. As shown in Table 3, our method outperformed all denoising methods. And our method can improve the SNR to 19.2 dB. This rate is close to 20 dB which the information extraction technologies have a good performance.

4.4 Subjective Evaluation

Second, subjective evaluation was carried out with 32 native English listeners. Twenty samples were randomly selected from the test set with SNRs of 2.5 and 7.5 dB. Listeners were presented two samples: noisy speech and denoised speech by each denoising method, as described in Table 3. The listener rated the overall quality of enhanced speech on a scale from 1 to 5, with a 1 being described as “BAD: degraded speech with very intrusive noise” and a 5 being “EXCELLENT: very natural speech with no degradation or noticeable noise”. Listeners could listen to each sample as many times as they wanted, and they were asked to focus on both signal distortion and noise intrusiveness.

Table 4 summarizes the result by averaging the MOS scores from all listeners. Our proposed method had the highest score compared with other methods.

4.5 Improvement of speech recognition

Finally, we evaluated the improvement of the speech recognition accuracy by speech denoising. Our intent here was to determine whether the denoised speech signals closely resemble the clean speech signals. As the accuracy of speech recognition depends highly on the acoustic model, we defined the recovery rate R from word error rate (WER) of the speech recognition as follows:

$$R = \frac{WER_{noisy} - WER_{denoised}}{WER_{noisy} - WER_{clean}}$$

where WER_{clean} , WER_{noisy} , and $WER_{denoised}$ mean WER in the case of using clean, noisy, and denoised

Table 4. Subjective MOS measures.

Method	MOS
Wiener	2.77
SEGAN	3.37
DFL	3.50
Proposal	3.90

Table 5. The result of recovery rate R by speech denoising.

Method	Julius	DeepSpeech
Wiener	0.13	-0.10
SEGAN	0.34	0.08
DFL	0.54	0.25
Proposal	0.63	0.41

speech signals respectively. If R is close to 1.0, the denoised speech signals are nearly equal to clean speech signals, and if R is less than 0.0, the speech denoising method worsens or does not affect the results. We used two of the open-to-the-public speech recognition engines, Julius⁽³⁰⁾ and DeepSpeech⁽³¹⁾. For Julius, we used version 4.4⁽³²⁾ and DNN English speech models for the acoustic model⁽³³⁾. For DeepSpeech, we used version 0.4.1⁽³⁴⁾ and the same version of English model. For each denoising model, we processed speech recognition from 834 noisy samples and calculated recovery rate from the recognized text.

Table 5 shows the result of recovery rate R by speech denoising when using Julius and DeepSpeech. As observed, all the speech denoising methods except “Wiener” label were effective in improving the accuracy of speech recognition, as the rates were more than 0. Comparing all methods, we can see that our proposed method outperformed “DFL”, which had the best performance among the conventional methods.

5. Conclusions

In this work, we applied a residual attention network on a U-Net architecture for end-to-end speech denoising. We evaluated both objective and subjective evaluations and found that our proposed method outperformed conventional methods. Future work will include further improvements such as adopting generative adversarial networks and more effective loss functions.

References

- (1) A. H. Moore, P. P. Parada and P. A. Naylor, “Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures”, Computer Speech & Language no. 46, pp. 574-584, Nov. 2017.
- (2) J. W. Kim, J. Salamon, P. Li and J. P. Bello, “CREPE: A Convolutional Representation for Pitch Estimation”, in Proc. IEEE ICASSP, Apr. 2018.

- (3) P. C. Loizou, *Speech Enhancement: Theory and Practice*, Second Edition. CRC Press, 2013.
- (4) N. Matsuo, C. Matsumoto, and S. Harada, "Super directional microphone array for vehicles", in *Proc. 19th ITS World Congress*, Oct. 2012.
- (5) F. Boll, Steven, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, 1979.
- (6) J. Lim and A. Oppenheim, "All-pole modeling of degraded speech", *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 3, pp. 197–210, 1978.
- (7) P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders", in *Proc. ICML*, pp. 1096–1103, July 2008.
- (8) X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder", in *Proc. INETRSPEECH*, pp. 436–440, Aug. 2013.
- (9) E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders", in *Proc. 5th IEEE Global Conference on Signal and Information Processing*, pp. 1265–1269, Nov. 2017.
- (10) D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation", *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, Mar. 2016.
- (11) S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network", in *Proc. INETRSPEECH*, pp. 3642–3646, Aug. 2017.
- (12) D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising", in *Proc. IEEE ICASSP*, Apr. 2018.
- (13) F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses", in *Proc. INETRSPEECH*, pp. 2723–2727, Sept. 2019.
- (14) O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation", in *Proc. MICCAI*, pp. 234–241, Oct. 2015.
- (15) F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions", in *Proc. ICLR*, May 2016.
- (16) P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image translation with conditional adversarial networks", in *Proc. CVPR*, pp. 1125–1134, July 2017.
- (17) Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net", *IEEE Trans. Geoscience and Remote Sensing, Lett.*, vol. 15, no. 5, pp. 749–753, 2018.
- (18) I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", in *Proc. NeurIPS*, pp. 2672–2680, Dec. 2014.
- (19) D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A multi-scale neural network for end-to-end audio source separation", in *Proc. ISMIR*, Sept. 2018.
- (20) A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio", *arXiv preprint arXiv:1609.03499*, Sept. 2016.
- (21) F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification", in *Proc. CVPR*, pp. 3156–3164, July 2017.
- (22) Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks", in *Proc. ECCV*, pp. 286–301, Sept. 2018.
- (23) K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proc. CVPR*, pp. 770–778, July 2016.
- (24) J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks", in *Proc. CVPR*, pp. 7132–7141, June 2018.
- (25) C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks", in *Proc. INETRSPEECH*, pp. 352–356, Sept. 2016.
- (26) <https://datashare.is.ed.ac.uk/handle/10283/2791>
- (27) <https://datashare.is.ed.ac.uk/handle/10283/2651>
- (28) <http://parole.loria.fr/DEMAND>
- (29) <https://www.tensorflow.org>
- (30) A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius", in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2009*, pp. 131–137, Oct. 2009.
- (31) A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates et al., "Deep Speech: Scaling up end-to-end speech recognition", *arXiv preprint arXiv:1412.5567*, Dec. 2014.
- (32) <http://julius.osdn.jp>
- (33) <https://sourceforge.net/projects/juliusmodels>
- (34) <https://github.com/mozilla/DeepSpeech>