ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10097114

HALF-TEMPORAL AND HALF-FREQUENCY ATTENTION U^{2}NET FOR SPEECH SIGNAL IMPROVEMENT

Zehua Zhang, Shiyun Xu, Xuyi Zhuang, YuKun Qian, Lianyu Zhou, Mingjiang Wang

Harbin Institute of Technology (Shenzhen), Shenzhen, China

ABSTRACT

During communication, volume changes, noise, and reverberation can disturb speech signals, significantly affecting the quality and intelligibility of speech. In the context of the ICASSP 2023 Signal Processing Grand Challenge, the first Speech Signal Improvement Grand Challenge (SIG) is organized to improve the quality of speech signals during communication. This paper proposes half-temporal and half-frequency attention U^{2}Net for improving full-band speech signal. Channel-spectrum attention is proposed for the skip connection between the encoder and decoder. The proposed model achieves 0.353, 1.289, 0.604, 0.625, and 0.924 improvements in signal, noise, overall, reverberation, and loudness, respectively, in the SIG subjective test. The proposed model achieved fourth place in the SIG real-time track, showing excellent denoising and de-reverberation performance.

***Index Terms*—** Speech signal improvement, speech enhancement, de-reverberation, temporal-frequency attention, U^{2}Net

1. INTRODUCTION

Noise and reverberation are the main disturbing components of the speech signal, and most speech enhancement methods focus on removing noise and reverberation. The first Speech Signal Improvement Grand Challenge[1] (SIG) [1] proposes an overall task to enhance the quality of speech signals and adds tasks to adjust loudness and improve coloration. Mainstream speech enhancement methods are in either the time or time-frequency domains. Considering that the features of speech signals in the time-frequency domain are more significant than those in the time domain and spectral mapping methods can be used to adjust loudness. This paper uses the complex spectrum as the learning target for deep neural networks. We propose a half-temporal and half-frequency attention U^{2}Net (HFHTAU^{2}Net) to improve speech signal. Half-temporal and half-frequency attention (HTHFA) and channel-spectrum attention (CSA) is proposed further to improve the model's attention to important features.

Thanks to the National Natural Science Foundation of China under Grant No. 62276076, the National Natural Science Foundation of China under Grant No. 62176102 and the Natural Science Foundation of Guangdong Province under Grant No. 2020B1515120004.

[1]https://www.microsoft.com/en-us/research/academic-program/speech-signal-improvement-challenge-icassp-2023/

2. PROPOSED METHODS

Our previous research [2] shows that extracting the full-band speech signal into three sub-band speech signals can effectively reduce the feature dimensionality. A similar structure is used in this paper, as shown in Fig. 1. We use U^{2}Net as the backbone network whose specific structure is similar to that paper [2]. HTHFA is added to each layer of the encoder and decoder to enhance the feature extraction ability. The skip connection between the encoder and decoder is replaced by the channel-spectrum attention module. The complex spectrum, magnitude spectrum, and phase features extracted by the phase encoder (PE) [3] are fed into the HTHFAU^{2}Net. The decoder (DC) includes a 2-D convolution and a linear layer to output the enhanced sub-band complex spectrum. The enhanced full-band speech signal is interpolated from three sub-band speech signals.

2.1. Half-Temporal and Half-Frequency Attention Module

Input tensors are divided into two parts in the channel dimension, one of which is temporal attention (T-attention), and the other is frequency attention (F-attention). Input tensors generate queries, keys, and values by point-wise convolution, batch normalization, and PRelu activation. The formulas for calculating F-attention and T-attention are shown in Eq. 1 and Eq. 2, where t, f, and C represents the frame index, the frequency bin, and the number of channels of the input tensor. In the training stage, an upper triangular masking matrix is added to T-attention to avoid the model from seeing future frames, and a lower triangular masking matrix is added to make the model focus only on 200 frames. HTHFA outputs are concatenated in the channel dimension and produced by point-wise convolution.

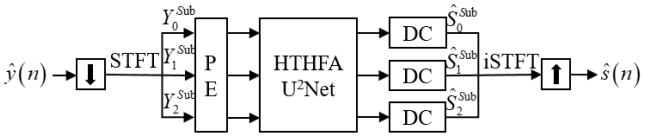

Fig. 1. Overall structure of the proposed HTHFAU^{2}Net

$$\mathbf{A}_F(t) = \text{Softmax}\left(\frac{\mathbf{Q}_f(t)\mathbf{K}_f'(t)}{\sqrt{C/2}}\right)\mathbf{V}_f(t) \quad (1)$$

$$\mathbf{A}_T(f) = \text{Softmax}\left(\frac{\text{Mask}\left(\mathbf{Q}_t(f)\mathbf{K}_t'(f)\right)}{\sqrt{C/2}}\right)\mathbf{V}_t(f) \quad (2)$$

2.2. Channel-Spectrum Attention Module

Inspired by the paper [4], the channel-spectrum attention module, as shown in Fig. 2, is proposed instead of the skip connection between the encoder and decoder. The average and maximum values of the input tensors are calculated in the frequency dimension and then added before a shared linear layer and Sigmoid activation to generate channel attention. Channel attention multiplied by the input tensor draws the model's attention to more key channels. The spectrum attention calculates the mean and maximum values in the channel dimension and then concatenates in the channel dimension and multiplies the input after 2-D convolution and Sigmoid activation.

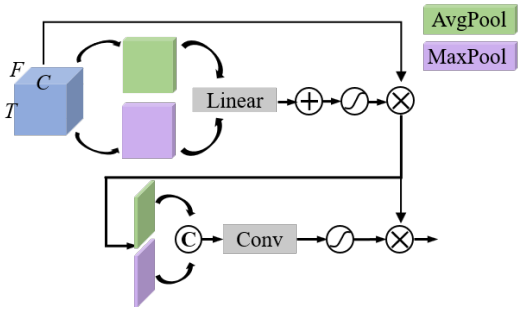

Fig. 2. The diagram of channel-spectrum attention module

3. EXPERIMENTS AND ANALYSIS

3.1. Dataset Construction

The clean speech and noise in the training dataset are from ICASSP 2022 DNS Challenge data [5]. We removed the poor-quality speech clips by DNSMOS [6]. We also recorded approximately 40 hours of real environment noise using voice recorders to extend noise data. The real reverberation is from the DNS [5], and synthetic reverberation is generated by the image method for RT60 from 0.4s to 1.4s. Clean speech has an eighty percent probability of being convolved with the reverberation and mixed with the noise at a random signal-to-noise ratio of -5dB to 15 dB. For the model to automatically adjust for loudness, thirty percent of the synthesized noisy speech is amplified, and thirty percent is reduced. Finally, we generated 1500 hours of clean-noisy speech pairs for the SIG and 300 hours for the ablation study.

Table 1. The subjective MOS evaluated on the SIG blind test set using P.804 (listening phase) / P.863.2 framework.

	Final Score	Signal	Noise	Overall
Noisy	0.411	2.927	3.302	2.360
Proposed	0.531	3.280	4.592	2.965
	Coloration	Discontinuty	Loudness	Reverberation
Noisy	3.029	4.061	2.992	3.852
Proposed	3.248	4.005	3.916	4.477

3.2. Experimental Result

On the SIG blind test set, the subjective evaluation results are shown in Table 1. Our model improved by 0.604, 0.218, 0.924, and 0.625 in overall, coloration, loudness, and reverberation, ranking fourth in the final score. Our model does not compensate for packet loss, so continuity is decreased. The proposed model has a real-time factor of 0.36 on the Intel Core i5 6400 CPU.

Table 2. Ablation study results in terms of DNSMOS

	Parameters	SIG	BAK	OVRL
Noisy	-	2.187	1.870	1.644
U2Net	8.947M	2.988	3.974	2.713
+PE	8.954M	3.026	3.966	2.738
+CSA	9.020M	3.041	3.996	2.767
+HTHFA	9.154M	3.227	4.056	2.944

U^{2}Net is used as the backbone network, and other modules are added successively. The ablation results are shown in the Tabel 2. PE improves overall performance at the cost of minimal model parameters. CSA only uses about 0.066M model parameters to improve performance compared to skip connection. HTHFA has the most significant improvement in model performance.

4. REFERENCES

[1] Ross Cutler, Ando Saabas, and et al., "Icassp 2023 speech signal improvement challenge," in *arXiv*, 2023.

[2] Z. Zhang and et al., "Fb-mstcn: A full-band single-channel speech enhancement method based on multi-scale temporal convolutional network," in *ICASSP*, 2022, pp. 9276–9280.

[3] G. Zhang and et al., "Multi-scale temporal frequency convolutional network with axial attention for multi-channel speech enhancement," in *ICASSP*, 2022, pp. 9206–9210.

[4] S. Woo and et al., "Cbam: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[5] H. Dubey and et al., "Icassp 2022 deep noise suppression challenge," in *ICASSP*, 2022.

[6] C. K. Reddy and et al., "Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*, 2022.