2021 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)

Convolutional Recurrent Neural Network With Attention Gates For Real-time Single-channel Speech Enhancement

Wen-Yu Wu
Dept. Communication Engineering
National Central University
Taoyuan, Taiwan
wywu.vaplab@gmail.com

Pin-Hsuan Li
Dept. Communication Engineering
National Central University
Taoyuan, Taiwan
phli.vaplab@gmail.com

Kai-Wen Liang
Dept. Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
kwistron@gmail.com

Pao-Chi Chang
Dept. Communication Engineering
National Central University
Taoyuan, Taiwan
pcchang@ce.ncu.edu.tw

***Abstract*—In this paper, we incorporate the attention gates (AG) into the convolutional recurrent neural network (CRNN) to perform speech enhancement. The attention gates, which enhance important features and suppress irrelevant parts, can help the system effectively generate more accurate complex ratio mask (CRM). Because the model takes into account the phase information, better speech quality can be obtained. Since the parameters of the proposed model can be reduced to only 2.3M, the computational complexity is low, and the objective of real-time speech enhancement can be achieved.**

Keywords—Deep Learning, Real-time Speech enhancement, Convolutional Recurrent Neural Network

I. INTRODUCTION

In today's indoor or outdoor environment, noises exist everywhere, which not only degrades the speech quality but also affects automatic speech recognition (ASR). Therefore, speech enhancement is a highly desired task when taking noisy speech as input and generates enhanced speech output to obtain better speech quality and intelligibility. Due to the popularity of deep learning (DL) technology, speech enhancement benefits from deep learning, which can effectively deal with non-stationary noise. In this paper, we focus on DL-based single-channel speech enhancement to obtain better perceptual quality and intelligibility, especially for real-time processing with low model complexity. The proposed system utilizes the convolutional recurrent neural network (CRNN) [1] with additional attention gates (AG) [2] to enhance important time frequency bands and suppress irrelevant parts to achieve better speech quality.

II. PROPOSED METHOD

In this paper, the flow diagram of the proposed system is shown as in Fig. 1. The proposed system mainly includes three parts, the preprocessing module, the neural network, and the loss function.

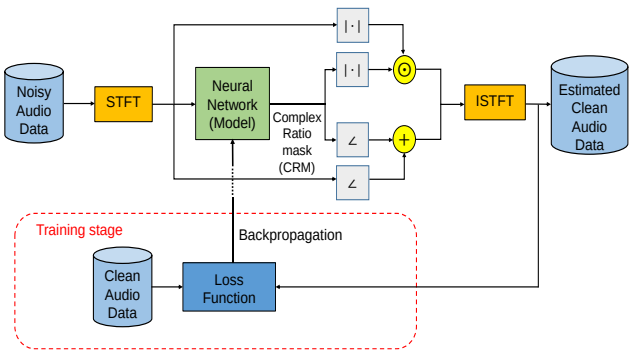

Fig. 1. The flow diagram of the proposed system.

978-1-6654-1951-2/21/$31.00 © 2021 IEEE

The preprocessing module mainly performs the short-time Fourier transform (STFT). We assume the microphone signals to be described in the STFT domain by (1)

$$X[k,l] = S[k,l] + N[k,l] \tag{1}$$

where $X[k,l]$, $S[k,l]$, and $N[k,l]$ denote the STFT at time frame l and frequency bin k of the noisy speech, clean speech, and noise, respectively. In our system, STFT is computed based on a 25 ms Hanning window with 75% overlap between frames and a 512-point discrete Fourier transform. STFT can be further decomposed into real and imaginary parts (2) (3), which are used as the input features of the neural network:

$$Real(X) = |X[k,l]|cos(\varphi_X) \tag{2}$$

$$Imag(X) = |X[k,l]|sin(\varphi_X) \tag{3}$$

where $|X[k,l]|$ and φ_X denote the magnitude and the phase of the noisy spectrogram, respectively.

The neural network model of the proposed system is shown as in Fig. 2. It is mainly composed of four parts, the encoder, the enhancer, the attention gates, and the decoder.

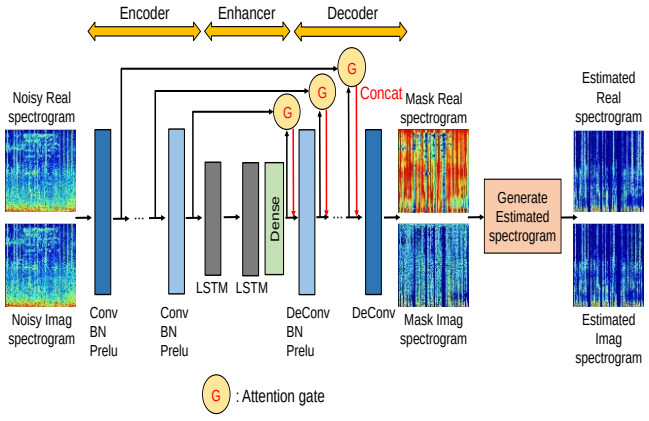

Fig. 2. The neural network model of the proposed system.

The encoder aims at extracting high-level features from the input features, and reducing the resolution. The input features are extracted through the five convolutional layers, each followed by batch normalization (BN) and Prelu activation. In the enhancer, the two long short-term memory (LSTM) layers are specifically used to model the temporal dependencies, and the one dense layer linearly adjusts the

features after LSTM. The AG focuses on the important time frequency bands and ignores the irrelevant parts to enhance the target speech. In the AG, it uses each layer of the encoder output features and the corresponding layer of the decoder input features, through 1×1 kernels, to get attention coefficients, which will be multiplied with the encoder output features. The three 1×1 kernels in an AG have the same number of channels, which are set by the current number of channels of convolutional layers, and BN is used after each convolutional operation. The flow diagram of the AG is shown as in Fig. 3. The decoder aims at reconstructing the low resolution features to the original size of input, leading the encoder-decoder structure to a symmetric design. The encoder output features after AG concatenate with the decoder input features, and go through the deconvolutional layers, BN, and Prelu activation. By adding AG, the decoder can effectively generate more accurate CRM.

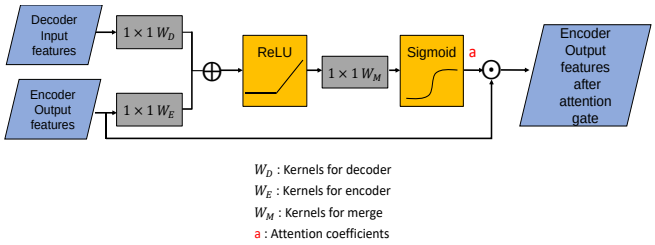

Fig. 3. The flow diagram of the attention gate.

The neural network generates the CRM as the output of the model. We can use the magnitude (4) and the phase (5) of the CRM to get the final estimated speech, represented by the magnitude (6) and the phase (7),

$$\left|\widetilde{M}\right| = tanh\left(\sqrt{\widetilde{M_r}^2 + \widetilde{M_i}^2}\right), 0 \le \left|\widetilde{M}\right| \le 1 \quad (4)$$

$$\angle\widetilde{M} = tan^{-1}\left(\frac{\widetilde{M_i}}{\widetilde{M_r}}\right) \quad (5)$$

$$\left|\tilde{S}\right| = |X| \odot \left|\widetilde{M}\right| \quad (6)$$

$$\angle\tilde{S} = \angle X + \angle\widetilde{M} \quad (7)$$

where $\widetilde{M}$ and $\tilde{S}$ denote the estimated mask and the estimated speech spectrogram, respectively. The loss function in our training stage is SI-SNR loss (8), which has been commonly used to replace the mean square error (MSE) loss [3]:

$$\begin{cases} s_{target} = (<\tilde{s}, s>\cdot s)/\|s\|_2^2 \\ e_{noise} = \tilde{s} - s_{target} \\ L_{SI-SNR} = -10\log_{10}\left(\frac{\|s_{target}\|_2^2}{\|e_{noise}\|_2^2}\right) \end{cases} \quad (8)$$

where s and $\tilde{s}$ denote the clean and estimated time-domain waveform, respectively, $<\cdot,\cdot>$ and $\|\cdot\|_2^2$ denote the dot product between two vectors and Euclidean norm (L2 norm).

III. EXPERIMENTAL RESULTS

In this work, the Deep Noise Suppression (DNS) Challenge 2020 datasets were used for our experiments. We set the audio length to 3 seconds, the sampling rate to 16 KHz, and the training and validation ratio to 95:5. For the total of 250 hours dataset, half of which were non-reverberant and the other half were reverberant. For testing our model, the results of PESQ and STOI of the test set were compared with other models.

TABLE I. PESQ and STOI on DNS challenge test set (simulated data)

Model	#Para. (M)	Process time (s)	No Reverb		Reverb	
			PESQ (MOS)	STOI (%)	PESQ (MOS)	STOI (%)
Noisy	-	-	2.454	91.52	2.753	86.62
NSNet [4] (Official)	5.1	-	2.873	94.47	3.076	90.43
DCCRN [3] (Baseline)	3.7	0.709	3.168	95.68	3.073	89.63
CRN	**2.2**	**0.322**	3.165	95.63	3.084	90.14
AGCRN (Proposed)	2.3	0.418	**3.213**	**95.86**	**3.147**	**90.37**

In TABLE I, we compared the official model and the baseline model. In DNS challenge 2020, DCCRN ranked first in the real-time speech enhancement task. The test results of the official model were referred to the reference [5], and the test results of the baseline model used the same 250 hours dataset for fair comparison. For the calculation of average processing time, we used a CPU to run the 3 seconds audio file. We can observe that using the previously introduced input features, neural networks, and the loss function, our proposed system has surpassed the official and the baseline models. Comparing with the baseline model, in terms of PESQ and STOI scores, PESQ improves by 0.06 MOS on average, while STOI improves by 0.46 % on average. In terms of the processing time and the amount of parameters, our proposed system also performs better than the baseline model. In addition, incorporating AG does help the neural network to effectively estimate the CRM and improve the final speech quality.

IV. CONCLUSION

In this work, we have proposed a system that includes a convolutional recurrent neural network with attention gates to estimate CRM. In the experimental results, our proposed model not only has surpassed the official and the baseline models on PESQ and STOI scores, but also performs better in processing time and the amount of parameters. By adding AG it does help the neural networks to effectively estimate the CRM. In the future work, we will try to put it in the edge device and to improve the performance under reverberation conditions.

REFERENCE

[1] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in Interspeech, vol. 2018, 2018, pp. 3229–3233.

[2] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz et al., "Attention u-net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999, 2018.

[3] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," arXiv preprint arXiv:2008.00264, 2020.

[4] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler and I. Tashev, "Weighted Speech Distortion Losses for Neural-Network-Based Real-Time Speech Enhancement," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 871-875.

[5] X. Hao, X. Su, R. Horaud and X. Li, "Fullsubnet: A Full-Band and Sub-Band Fusion Model for Real-Time Single-Channel Speech Enhancement," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6633-6637.