End-to-end speech-denoising deep neural network based on residual-attention gated linear units

Seon Man Kim✉
School of Computing and Artificial Intelligence, Hanshin University, Gyeonggi, 18101, South Korea
✉E-mail: smkim@hs.ac.kr, kobem30002@gmail.com

In this letter, an improved gated linear unit (GLU) structure for end-to-end (E2E) speech enhancement is proposed. In the U-Net structure, which is widely used as the foundational architecture for E2E deep neural network-based speech denoising, the input noisy speech signal undergoes multiple layers of encoding and is compressed into essential potential representative information at the bottleneck. The latent information is then transmitted to the decoder stage for the restoration of the target clean speech. Among these approaches, CleanUNet, a prominent state-of-the-art (SOTA) method, enhances temporal attention in latent space by employing multi-head self-attention. However, unlike the approach of applying the attention mechanism to the potentially compressed representative information of the bottleneck layer, the proposed method instead assigns the attention module to the GLU of each encoder/decoder block layer. The proposed method is validated by measuring short-term objective speech intelligibility and sound quality. The objective evaluation results indicated that the proposed method using residual-attention GLU outperformed existing methods using SOTA models such as FAIR-denoiser and CleanUNet across signal-to-noise ratios ranging from 0 to 15 dB.

Introduction: Speech denoising plays a pivotal role in improving the overall user experience by minimizing environmental interference and promoting a better understanding of speech content in the field of speech-based information technology services such as automatic speech recognition, hearing aids, and communication systems. To achieve this goal, speech-denoising methods utilizing deep neural networks (DNNs) have been extensively researched over the last decade [1], and it is widely acknowledged that these DNNs can substantially enhance speech intelligibility and quality compared to conventional methods, particularly within the domain of speech processing.

In the field of DNNs, the U-shaped encoder-decoder structure, known as UNet, is predominantly used for speech enhancement, with the speech spectrum typically serving as the input feature. This approach involves estimating the clean speech spectrum from the noisy speech spectrum [1, 2]. To maintain temporal consistency of speech features within this UNet framework, recurrent neural networks (RNNs), such as long short-term memory (LSTM), are often incorporated at the bottleneck between the encoder and decoder [3]. Additionally, attention mechanisms have been introduced to further improve temporal consistency at the bottleneck [4]. Moreover, it has recently become important to develop models that restore the original speech phase information while considering only the spectral magnitude. To address this, two strategies are being actively researched: using complex NN modules [3, 5] to predict the complex domain spectrum of speech, and employing an end-to-end (E2E) architecture that aims to directly predict target signal waveforms from existing waveform inputs. Among these two main approaches, this letter focuses on speech denoising based on the E2E architecture.

A representative example of the E2E architecture is Wave-U-Net [6], a U-shaped architecture that incorporates skip connections between encoders and decoders through multiple Conv1D layers. This model is currently used as the foundational architecture for several speech-denoising approaches. In the Wave-U-Net structure, the input noisy speech signal is compressed into latent representative information at the bottleneck by several encoding layers; this latent information is gradually restored to the target clean speech using decoding layers [6]. To further improve the speech-denoising performance of the Wave-U-Net architecture, recurrent neural networks such as LSTM have been applied to maintain the temporal context for potential representative information in the bottleneck layer [7]. Furthermore, in [8], the temporal context was strengthened by applying multi-head self-attention (MHA) to the potential information of the bottleneck layer. The representative model utilizing LSTM in the bottleneck is FAIR-denoiser, and the representative model utilizing MHA is CleanUNet, with CleanUNet showing better performance than FAIR-denoiser [8].

Meanwhile, the methodology for enhancing the temporal context of a bottleneck assumes that the potentially representative information of the bottleneck is already well compressed by the encoding stage. Therefore, if the encoder layers gradually compress the potential representative information toward the bottleneck, the target clean speech can be expected to be well restored by the bottleneck and decoder layers. Taking this into consideration, we aim to enhance speech-denoising performance by applying attention functions to the encoder and decoder layers before and after the bottleneck. To this end, we specifically propose a residual-attention GLU (RAGLU) that integrates a GLU and a residual-attention module applied to each encoder/decoder layer. Additionally, we propose a new E2E architecture based on RAGLU.

Proposed RAGLU methodology: Before explaining RAGLU, the original GLU will be briefly explained. The basic architecture of GLU is shown in Figure 1a. GLU doubles the feature extraction of the convolutional neural network, half of which is used as a gating signal by the sigmoid function to control the flow of information in the other half [9]. With this mechanism, GLU helps the model focus on relevant information for a given task by selectively emphasizing or weakening different aspects of the input features. GLUs are used in all layers of the encoder and decoder within E2E speech-denoising architectures such as FAIR-denoiser and CleanUNet, helping to identify event-related information in the time–frequency domain and filter out irrelevant background noise [7, 8].

In this work, I did not intend to identify or solve particular problems or weaknesses of GLU. Rather, I attempted to further enhance the speech-denoising efficiency of GLU. Among the features extracted within GLU that were doubled in size, excluding half the features used as the gating signal using a sigmoid function, I wanted to further strengthen the speech feature map of the remaining half used as the main signal. To this end, the proposed RAGLU sought to strengthen the channel and temporal context within the signal using the residual-attention network for noise-robust feature map refinement in the main signal part inside GLU. In detail, focusing on the structural diagram of the proposed RAGLU in Figure 1b, channel and temporal attention based on the convolutional block attention module (CBAM) [10] are applied to the first half of the features extracted by a 1×1 convolutional neural network. CBAM is introduced to focus on the most salient parts of specific features rather than the entire scene. In particular, compared to other attention-based methods, CBAM has fewer parameters and maintains high speed. These channel and temporal attentions were implemented following an existing methodology [10], and the description of CBAM is omitted in this letter. For further details, please refer to [10]. Notably, the CBAM methodology in [10] targets a two-dimensional feature map; therefore, the term spatial attention was used. However, the wave signal for E2E speech denoising covered in this work corresponds to a feature map in the one-dimensional temporal domain. It corresponds to the concept of temporal attention. Thus far, I have explained RAGLU for speech denoising. It is important to note that the proposed RAGLU is only one activation unit, and speech-enhancing performance improvement can be achieved simply by replacing the GLU of the existing E2E speech-denoising model with RAGLU. The verification of this process is presented in the Results and Discussion section.

Below, I present an E2E speech-denoising architecture designed to effectively utilize the proposed RAGLU. Figure 2 illustrates the architecture of the proposed model. The primary aim is to replace the GLU utilized in E2E speech-denoising architectures such as FAIR-denoiser and CleanUNet with the proposed RAGLU. This approach emphasizes speech feature compression at the encoder–decoder stage before and after the bottleneck. Consequently, the bottleneck does not introduce an attention mechanism with a large model size such as MHA; rather, it simply maintains the temporal context of speech features using a sequential network such as LSTM. In particular, the proposed model comprises eight blocks of encoder–decoder pairs and a bottleneck, similar to CleanUNet and FAIR-denoiser. The first four blocks constitute the encoder–decoder stage, while the subsequent four blocks form the attention encoder–decoder stage. The number of encoder–decoder block channels starts from 64; is sequentially doubled to 128, 256, and 512;

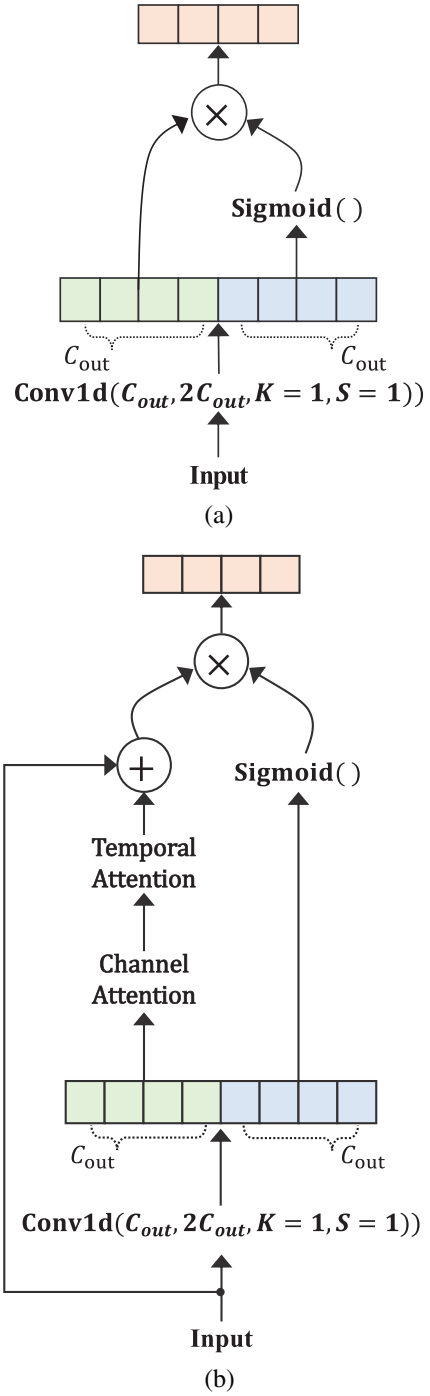

Fig. 1 *Comparison between the proposed residual attention gated linear unit (RAGLU) and the original GLU. (a) Original GLU architecture. (b) Proposed RAGLU architecture*

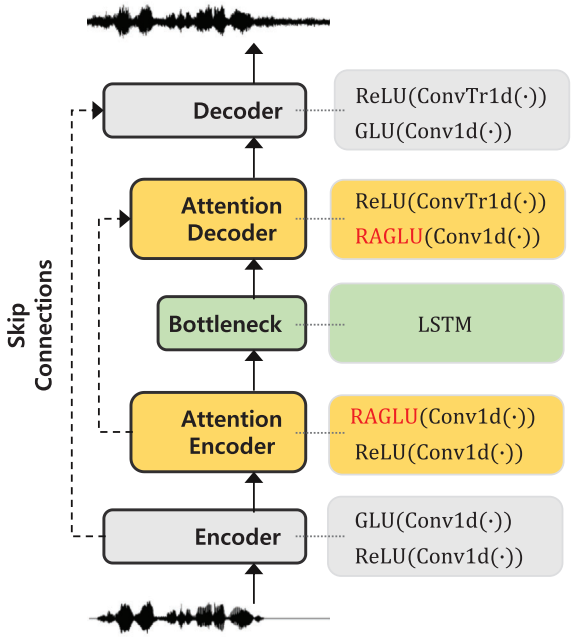

Fig. 2 *Attention encoder–decoder stage based causal end-to-end speech-denoising architecture using the proposed residual attention gated linear unit*

and is fixed to 768 in the subsequent four attention encoder–decoder blocks.

Additionally, in the encoder stage, each block is sequentially executed by Relu(Conv1d(C_{in}, C_{out}, $K = 8$, $S = 4$)) and GLU(Conv1d(C_{out}, $2C_{out}$, $K = 1$, $S = 1$)). In the decoder stage, the blocks GLU(Conv1d(C_{in}, $2C_{in}$, $K = 1$, $S = 1$)) and Relu(ConvTr1d(C_{in}, C_{out}, $K = 8$, $S = 4$)) are processed sequentially. Here, C_{in}, C_{out}, K, and S represent the input channel size, output channel size, kernel size, and stride size, respectively. In the attention encoder–decoder stage before and after the bottleneck stage, RAGLU is applied instead of GLU to each block. Specifically, in the attention encoder, Relu(Conv1d(C_{in},C_{out}, $K = 4$, $S = 2$)) and RAGLU(Conv1d(C_{out},$2C_{out}$, $K = 1$, $S = 1$)) are processed sequentially in each block. In the attention decoder, RAGLU(Conv1d(C_{in},$2C_{in}$, $K = 1$, $S = 1$))and Relu(ConvTr1d(C_{in}, C_{out}, $K = 4$, $S = 2$)) are processed sequentially. Next, at the bottleneck, an LSTM comprising two layers and 768 hidden units is applied following the methodology in [7]. For causal processing, a unidirectional LSTM is used instead of a bidirectional LSTM.

Experimental setup: An experiment was conducted to evaluate the speech-denoising performance of RAGLU by leveraging the Valentini dataset available in the Voice Bank corpus as in [11]. The dataset comprises both clean and noisy speech samples recorded at a sampling rate of 48 kHz from 30 speakers. For the experiment, all samples were resampled to 16 kHz. The Valentini DB is specifically designed to address performance in domain mismatch scenarios by employing different speakers and noise types during both the training and validation stages. Specifically, 28 of the 30 speakers were allocated to the training dataset, while the remaining two were designated for the testing dataset. The noisy conditions cover a range of ten distinct noise types, comprising two synthetic noises (speech-shaped noise and babble noise) and eight authentic environmental noises sourced from the DEMAND database [11]. The real-world noises are categorized into domestic kitchen settings; office meeting room environments; public spaces such as cafeterias, restaurants, and subway stations; transportation settings such as cars and metro trains; and urban settings such as busy intersections. Accordingly, the training dataset comprised 11,572 pairs of clean and noisy speech segments across four signal-to-noise ratio (SNR) levels: 0, 5, 10, and 15 dB. Conversely, the test dataset comprised 824 pairs of clean and noisy speech samples, incorporating five distinct noise types (living rooms, office spaces, buses, outdoor cafeterias, and public squares) at SNRs of 2.5, 7.5, 12.5, and 17.5 dB, sourced from the DEMAND database, ensuring variability from the training dataset [11].

This experiment focuses on analysing the performance improvement effect within the E2E UNet architecture of RAGLU; for this purpose, FAIR-denoiser and CleanUNet are utilized as the basic architectures. To train a pair of noisy and clean signals, 4.5 s clips were randomly taken for both signals. Various data-augmentation techniques such as Remix, Band-mask, and Revecho were then applied to these clips. During the training process, the batch size was 64, the maximum epoch was set to 400, and the early-stopping parameter was set to 300. The Adam optimizer with momentum $\beta 1 = 0.9$ and denominator momentum $\beta 2 = 0.999$ was used. In addition, the linear warmup with the cosine annealing learning rate schedule with a maximum learning rate $= 2 \times 10^{-4}$ and warmup ratio $= 5\%$ was used. Finally, to update the network model during the learning process, full-band short-term Fourier transform loss was employed. All models were trained on six NVIDIA A40 GPUs.

Results and discussion: To assess the signal-denoising performance—the primary focus of this study—the perceptual evaluation of speech

Table 1. Comparison of objective evaluation results of short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) scores to validate the speech-enhancement benefits of residual attention gated linear unit (RAGLU) in CleanUNet

DNN model	STOI (%)	PESQ	Model size (MB)
CleanUNet	94.75	2.953	180.0
CleanUNet with RAGLU	94.95	3.030	190.7
CleanUNet with RAGLU without MHA	94.77	3.024	152.6

Abbreviations: DNN, deep neural network; MHA, multi-head self-attention.

Table 2. Comparison of objective evaluation results of short-time objective intelligibility (STOI) and perceptual evaluation of speech quality (PESQ) score enhancement by various methods

DNN models	STOI (%)	PESQ	Model size (MB)
DCCRN	93.39	2.913	11.53
Wave-U-Net	93.87	2.843	39.00
Denoiser	94.54	2.904	73.70
CleanUNet	94.75	2.953	180.0
PR	94.96	3.016	168.3

Abbreviations: DCCRN, deep complex convolution recurrent network; DNN, deep neural network; PR, proposed model.

quality (PESQ) [8] and short-time objective intelligibility (STOI) [8] scores, which are commonly employed in speech enhancement evaluation, were utilized. PESQ measures the effectiveness of noise cancellation in improving speech quality in the range of 0–4.5, with higher values indicating better speech quality. Additionally, STOI serves as an objective indicator of how well the target speech component is recovered in a noisy environment, ranging from 0 to 1, with larger values indicating greater recognition.

First, to evaluate the speech-denoising performance with the proposed RAGLU, it was integrated into CleanUNet—a current state-of-the-art model—in two scenarios. In the first scenario, all GLUs in CleanUNet's encoder and decoder blocks were replaced with RAGLU. This allows us to assess the advantages of RAGLU over existing GLUs in terms of computational efficiency and denoising performance. In the second scenario, RAGLU was applied after excluding MHA, which is included in the bottleneck of the existing CleanUNet. This scenario helps determine whether RAGLU can replace MHA for speech denoising and, if so, to what extent it improves speech denoising performance and computational efficiency. The results presented in Table 1 show that in the first scenario—where CleanUNet's GLUs are replaced with RAGLU—the model size increases by 10.7 MB, but STOI and PESQ values improve significantly. In the second scenario, applying RAGLU without MHA achieves similar or even better performance while making the model approximately 38 MB lighter than the baseline CleanUNet with MHA. In conclusion, the proposed RAGLU not only enhances speech denoising performance but also reduces model size sufficiently, making it a viable replacement for MHA in E2E speech denoising models like CleanUNet.

The results of PESQ, STOI, and model size for existing speech denoising models, including DCCRN [5], Wave-U-Net [6], Denoiser [7], and CleanUNet [8], as well as the proposed model (PR) depicted in Figure 1, are summarized in Table 2. As revealed by Table 2, the proposed model demonstrated significant performance improvements in terms of both STOI and PESQ compared to existing methods. Notably, the proposed model enhanced speech denoising performance in terms of STOI and PESQ while reducing the model size to approximately 11 MB compared to CleanUNet. Additionally, when comparing the proposed model to 'CleanUNet with RAGLU' in Table 1, which demonstrates the best speech denoising performance among the scenarios considered in this study, the proposed model maintains equivalent performance in terms of STOI and PESQ while having a model size that is 22 MB smaller.

Conclusion: In this letter, I propose RAGLU, which integrates the attention function with the commonly used GLU within the U-Net architecture, serving as a fundamental framework for E2E DNN-based speech denoising. Unlike existing methods such as MHA, which apply attention mechanisms to the representative information of the potentially compressed bottleneck layer, the proposed method involves incorporating an attention mechanism into each encoder/decoder RAGLU block. The proposed method demonstrated superior performance compared to existing methods such as FAIR-denoiser and CleanUNet, particularly in terms of objective measures of speech intelligibility and quality.

Acknowledgements: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2022R1A2C2010614).

Conflict of interest statement: The authors declare no conflicts of interest

Data availability statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received: *31 March 2024* Accepted: *27 August 2024*
doi: 10.1049/ell2.70020

References

1 Ali, J., Saleem, N., Bourouis, S., Alabdulkreem, E., Mannai, H.E., Dhahbi, S.: Spatio-temporal features representation using recurrent capsules for monaural speech enhancement. *IEEE Access* **12**, 21287–21303 (2024). https://doi.org/10.1109/ACCESS.2024.3361286
2 Khattak, M.I., Saleem, N., Gao, J., Verdu, E., Fuente, J.P.: Regularized sparse features for noisy speech enhancement using deep neural networks. *Comput. Electr. Eng.* **100**, 107887 (2022). https://doi.org/10.1016/j.compeleceng.2022.107887
3 Wahab, F.E., Ye, Z., Saleem, N., Ullah, R.: Compact deep neural networks for real-time speech enhancement on resource-limited devices. *Speech Commun*. **156**, 1–11 (2024). https://doi.org/10.1016/j.specom.2023.103008
4 Peracha F.K., Khattak M.I., Salem N., Saleem N.: Causal speech enhancement using dynamical-weighted loss and attention encoder-decoder recurrent neural network. *PLoS One* **18**, 1–22 (2023). https://doi.org/10.1371/journal.pone.0285629
5 Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., Xie, L.: DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. *arXiv:2008.00264* (2020). https://doi.org/10.48550/arXiv.2008.00264
6 Gan, X., Zheng, Z., Zeng, Q.: Speech enhancement algorithm based on Wave-U-Net. In: Proceedings of the ISCTIS 2023—3rd International Symposium on Computer Technology and Information Science, pp. 433–437. IEEE, Piscataway (2023)
7 Defossez, A., Synnaeve, G., Adi, Y.: Real time speech enhancement in the waveform domain. In: Proceedings of the Interspeech 2020, China, pp. 3291–3295. International Speech Communication Association, San Francisco (2020)
8 Kong, Z., Ping, W., Dantrey, A., Catanzaro, B.: Speech denoising in the waveform domain with self-attention. In: Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7867–7871. IEEE, Piscataway (2022)
9 Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: Proceedings of the ICML 2017, pp. 933–941. Microtome Publishing, Brookline, MA (2017)
10 Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision, pp. 3–19. Springer, Cham (2018)
11 Valentini-Botinhao, C., Wang, X., Takaki, S., Yamagishi, J.: Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks. In: Proceedings of the Interspeech 2016, pp. 352–356. International Speech Communication Association, San Francisco (2016)