NCSP'25 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing 2025
February 27 - March 2, 2025, Pulau Pinang, Malaysia

Global Convolutional-Block-Attention-Module-based Diffusion Model for Speech Enhancement

Yang Li, Zheng Qiu and Shoji Makino

Waseda University
2-7 Hibikino, Wakamatsu-ku, Kitakyushu, Fukuoka, Japan
E-mail: {itsluy@akane., qiuzheng@akane., s.makino@}waseda.jp

Abstract

Speech enhancement aims to extract or recover clean speech from noisy audio signals. Many recent studies employ diffusion-based generative models for this task. Score-Based Generative Model for Speech Enhancement (SGMSE) uses a U-Net-based score model to estimate parameters in the reverse process. This approach requires significant computational resources and large datasets for training and processing. In this paper, we propose the Global Convolutional Block Attention Module (Global CBAM), a lightweight model for training on small datasets. Our method enhances the global channel and spatial information of feature maps and effectively boosting denoising and speech enhancement capabilities. Besides, we employ a joint loss function to support model training. Simulation results show that the proposed method outperforms baseline models on both large and small datasets.

1. Introduction

Speech enhancement aims to restore noisy speech signals to clean speech. Diffusion generative models have recently gained attention in the field of speech enhancement and have shown good results [1, 2, 3, 4]. It is usually used with a network model such as U-Net. Machine learning algorithms can be used to extract these statistical properties by learning useful representations from large datasets [1]. However, training and processing large datasets require significant time and often result in poor performance on smaller datasets. Therefore, it is important to find ways to efficiently train models on small datasets and process complex noisy speech.

Recently, score-based generative models have proven effective in the field of audio enhancement. However, the performance of the model is limited when working with small training datasets due to insufficient parameter learning and a lack of data diversity. At the same time, the model cannot accurately identify and enhance the speech signal in the face of complex noise conditions. The challenge is to improve the results without consuming additional training resources.

The output of a predictive model has been proposed as the input for a diffusion model [2, 5, 6]. The combination of predictive and generative models uses the predictive model's ability to enhance speech and the generative model's ability to refine details, achieving efficient and accurate speech enhancement. During the speech generation process, the predictive model simplifies the task and reduces the computational load of the diffusion model by generating a preliminary enhanced speech signal that serves as input for the reverse process. This approach not only accelerates the generation process but also mitigates the potential artifacts introduced by the generative model. However, it significantly increases both the training time due to the need to train two models. Therefore, in this paper, we try to explore structural optimizations within the model itself to address these challenges.

Extensive research has been conducted to develop more effective methods for processing audio in complex acoustic environments. One approach [6] involves the use of predictive networks to preprocess noisy speech. This method significantly improves the final enhancement results and enhances the model's robustness by preprocessing the input signal. However, it requires substantial training resources, involves complex processing steps, and lacks simplicity in model design. Another approach [1] aims to improve performance by optimizing the model structure with advanced network designs. Although this simplifies computation, it often results in overly large networks for complex audio conditions.

In this paper, we propose a Global Convolutional Block Attention Module (Global CBAM), which is based on the Convolutional Block Attention Module (CBAM). [7]. This is a lightweight and efficient module that makes the model not only computationally efficient but also more effective for speech enhancement tasks. Besides, a joint loss function is employed to improve the model's performance, enabling better handling of complex noise conditions. Here, we use the Score-Based Generative Model for Speech Enhancement (SGMSE) as the mathematical framework for modeling. We then designed a Global CBAM and integrated it into the score model. This helps the network better capture global information from the input features. We performed comparative experiments, and the results indicate that our proposed method outperforms SGMSE on both large and small datasets.

2. Conventional diffusion model based speech enhancement

We use the Score-Based Generative Model for Speech Enhancement (SGMSE) [1] as the baseline in this article.

To achieve the speech enhancement task, they define a forward process and a reverse process as shown in Figure 1. Audio processing is performed in the frequency domain, where the real and imaginary parts of the complex signal are treated as two separate channels. In the forward process, the clean speech signal $\mathbf{x}_0$ is gradually transformed into the noisy speech signal $\mathbf{x}_T$. The variable $t \in [0, T]$ represents a continuous time-step that indicates the progression of the process over its duration. By step T, clean speech has been converted into noisy speech. In the reverse process, the noisy speech signal $\mathbf{x}_T$ is gradually restored to $\mathbf{x}_0$ using the score s_θ generated by the score model. The forward process as follows is a stochastic diffusion process $\{\mathbf{x}_t\}_{t=0}^{T}$ that is modeled as the solution to a linear Stochastic Differential Equation (SDE) of the general form [1],

$$\mathrm{d}\mathbf{x}_t = \gamma\left(\mathbf{y} - \mathbf{x}_t\right)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \tag{1}$$

where $\mathbf{x}_t$ is the present state, and $\mathbf{y}$ is the speech signal with noise, $\gamma(\mathbf{y} - \mathbf{x}_t)$ is the drift coefficient, which controls the gradual evolution of the signal towards the target noise distribution, γ is a constant that controls the rate of noise addition, $\mathbf{w}$ represents the Wiener process, and $g(t)$ is the diffusion coefficient, which controls the amount of Gaussian noise added at each time t, which is defined as:

$$g(t) := \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{t}\sqrt{2\log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}, \tag{2}$$

where $\sigma_{\min}$ and $\sigma_{\max}$ are parameters defining the noise schedule of the Wiener process. The reverse process of SGMSE is realized by solving the following SDE,

$$\mathrm{d}\mathbf{x}_t = \left[-\gamma\mathbf{f}\left(\mathbf{x}_t, \mathbf{y}\right) + g(t)^2\nabla_{\mathbf{x}_t} p_t\left(\mathbf{x}_t|\mathbf{y}\right)\right]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}. \tag{3}$$

The key part of the reverse process is computing the score $\nabla_{\mathbf{x}_t} p_t\left(\mathbf{x}_t|\mathbf{y}\right)$ generated by the score model, which is represented as $s_\theta(\mathbf{x}_t, \mathbf{y}, t)$, used to estimate the gradient of the conditional distribution $\log p_t(\mathbf{x}_t \mid \mathbf{y})$ with respect to $\mathbf{x}_t$. Here, θ denotes the set of parameters, covering all weights and biases in the network, used for training and optimization. The loss function used here is the mean squared error (MSE).

The score model here is used to estimate the gradient of the noise distribution, with its output being the score $\nabla_{\mathbf{x}_t} p_t\left(\mathbf{x}_t|\mathbf{y}\right)$. The main parts of the score model use the Noise Conditional Score Network (NCSN++) [8]. This is a Multi-Resolution U-Net Structure, consisting of upsampling layers, downsampling layers, progressive upsampling layers, progressive downsampling layers, and a bottleneck layer. As shown in Figure 2, the encoder and decoder correspond to the downsampling and upsampling layers, respectively. Except for the bottleneck layer, each part comprises six layers. The upsampling and downsampling layers are based on residual network blocks derived from the BigGAN architecture [9]. Each upsampling layer includes three residual blocks, while each downsampling layer contains two residual blocks, with

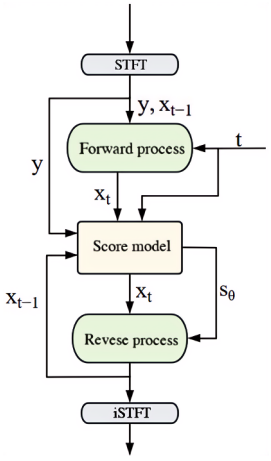

Figure 1: Flowchart of the Diffusion Model

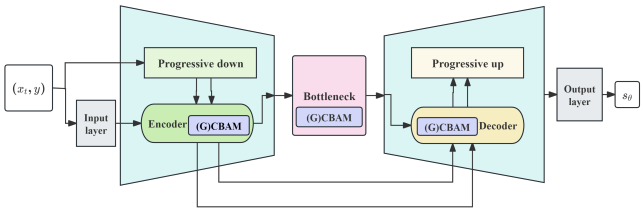

Figure 2: Structure of the score model

the final residual block performing the upsampling or downsampling operation.

Since the score model is based on the U-Net architecture, a skip connection is performed between each corresponding downsampling and upsampling layer, where the feature values are transferred via concatenation. This design aims to preserve high-resolution features and combine them with low-resolution decoding features to recover details. Additionally, the network incorporates a progressively growing input, as illustrated by the progressive down and up layers in Figure 2. The purpose is to provide a downsampled version of the input to each feature map in the contracting path, which has been successfully used to stabilize high-resolution image generation [10].

However, SGMSE performs ineffectively when dealing with complex data with limited training datasets. Moreover, it only considers spatial information while ignoring channel information.

3. Proposed method

3.1 Joint loss function

To enhance the model's performance in handling complex noise signals, we decide to use the following equation as the training objective,based on a combination of Mean Squared

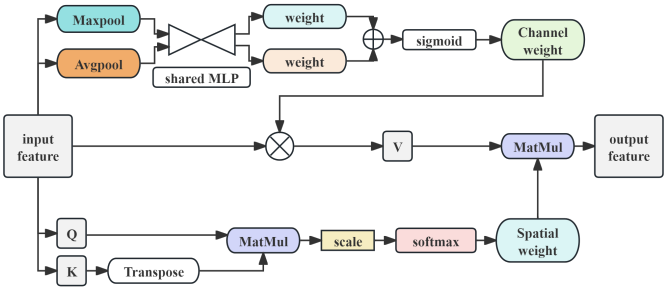

Figure 3: Structure of proposed Global CBAM

Error (MSE) and Mean Absolute Error (MAE), which is

$$\arg\min_{\theta} \mathbb{E}_{t,(\mathbf{x}_0,\mathbf{y}),z,\mathbf{x}_t|(\mathbf{x}_0,\mathbf{y})} \left[\left\| s_\theta(\mathbf{x}_t,\mathbf{y},t) + \frac{z}{\sigma(t)} \right\|_2^2 + \alpha \cdot \left\| s_\theta(\mathbf{x}_t,\mathbf{y},t) + \frac{z}{\sigma(t)} \right\|_1 \right]. \quad (4)$$

as shown in Equation 4, defined as MSE + αMAE, where α is a hyperparameter. We decided to use the precision of MSE as the foundation while introducing αMAE to better handle complex noise signals. We did the ablation experiment and demonstrate that this proposed method is effective.

3.2 Global Convolutional Block Attention Module

Figure 3 shows the structure of our proposed Global Convolutional Block Attention Module (Global CBAM). Similarly to Convolutional Block Attention Module (CBAM) [7], it consists of a channel attention module and a spatial attention module. In our score model, we added the proposed Global CBAM at the 16×16 resolution and in the bottleneck layer. This is to make better use of the channel attention for feature enhancement and to improve the model's understanding of global spatial information in the feature maps. The other parts of the score model use the Noise Conditional Score Network (NCSN++) [8].

The part of the channel attention we follow the structure of CBAM. Then we use the global spatial attention [11] module to replace the local spatial convolutional attention module. This design enables our proposed method to enhance both global spatial and channel information, providing stronger adaptability when dealing with complex datasets.

4. Experiment

4.1 Dataset

We generated the CHiME3-WSJ0 dataset, which was produced by cleaning speech from the Wall Street Journal (WSJ0) dataset [12] and noise signals from the CHiME3 dataset[13]. In the generated dataset, some extremely noisy segments are included to test the robustness of the model. The dataset contains 12777 speech files.We selected 1000 samples from the training set as a small dataset. We consider the entire dataset as the large dataset. The model was trained for 30 epochs on the entire dataset and 15 epochs on the small dataset.

4.2 Evaluation metrics

We used Perceptual Evaluation of Speech Quality (PESQ) [14], Extended Short-Time Objective Intelligibility (ESTOI) [15], Signal-to-Distortion Ratio (SDR) [16], and Signal-to-Noise Ratio Improvement (SNRi) as evaluation metrics, aiming to comprehensively assess the quality of the generated enhanced speech. Using SNRi is because the SNR improvement is more pronounced, thus providing a greater reference value. All metrics were calculated using the same methods as those in SGMSE, ensuring the reliability of the experiments.

4.3 Hyperparameters and Training Configuration

To investigate the influence of different values of α on the results, we conducted a set of experiments by replacing the baseline's original MSE loss function with our proposed MSE + αMAE, where α ranged from 0.05 to 0.35. Compared with the baseline that employs only MSE, the best performance was observed when $\alpha = 0.3$. Therefore, we set $\alpha = 0.3$ for all subsequent experiments.

The experiment process is illustrated in Figures 4 and Figures 5, where it can be observed that the model performance exceeds the baseline models in all metrics in this setting. The results of the experiment were obtained after training 15 epochs on small dataset with SI-SDR [17] and PESQ used as evaluation metrics.

The input and output layers of the score model are Conv2D layers with a 3×3 kernel and a stride of 1.

4.4 Experimental results

The results are shown in Table 1 and Table 2. Here, Proposed 1 refers to CBAM + joint loss, while Proposed 2 corresponds to Global CBAM + joint loss. The table shows that using CBAM with joint loss significantly improves model performance on small training datasets, and the proposed Global CBAM with joint loss further enhances performance on large datasets.

Figure 6 shows a comparison of spectrograms between the results of our proposed method and SGMSE. The results were obtained using a large dataset with the proposed Global CBAM and joint loss function. As shown in the white squares, our method demonstrates a more effective recovery of voice patterns, resulting in higher audio quality. The red squares highlight noise that the baseline method failed to remove accurately, which is effectively eliminated by our proposed method.

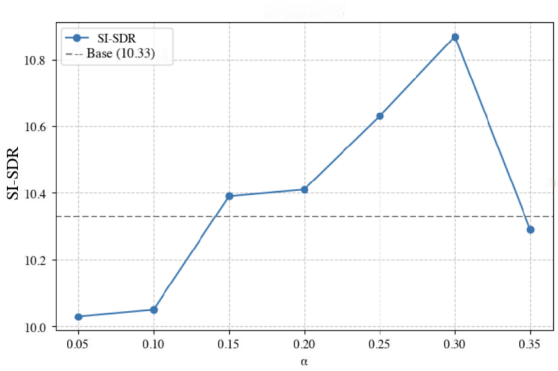

Figure 4: Experiment results on SI-SDR

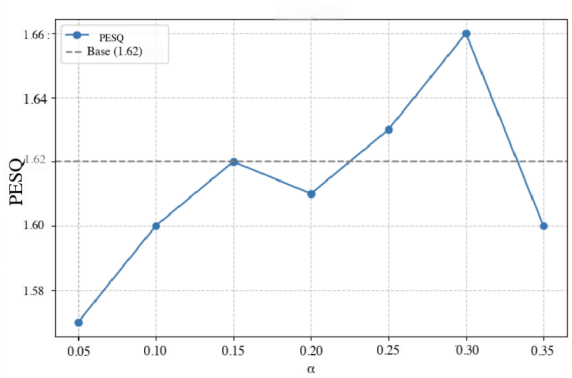

Figure 5: Experiment results on PESQ

5. Conclusions

In this paper, we propose a Global CBAM-based diffusion generative model for the speech enhancement task. Furthermore, we employ a joint loss function combining Mean Squared Error (MSE) and Weighted Mean Absolute Error (αMAE). This joint loss function enhances the model's ability to handle complex noise conditions. The Global CBAM strengthens feature representation, ensuring improved audio quality and enhancement performance. The results demonstrate that our model achieves performance improvements on both large and small datasets.

Table 1: Comparison of Speech Enhancement Results for SGMSE and Two Proposed Methods on Small Dataset (1000 Training Sets)

Method	PESQ	ESTOI	SDR	SNRi
SGMSE [1]	1.54	0.67	9.24	0.97
Proposed 1	**1.58**	**0.70**	**10.06**	**1.80**
Proposed 2	1.57	0.69	10.03	1.78

Table 2: Comparison of Speech Enhancement Results for SGMSE and Two Proposed Methods on Large Dataset

Method	PESQ	ESTOI	SDR	SNRi
SGMSE [1]	2.52	0.89	16.15	7.04
Proposed 1	2.63	0.89	16.20	7.07
Proposed 2	**2.67**	**0.90**	**16.60**	**7.48**

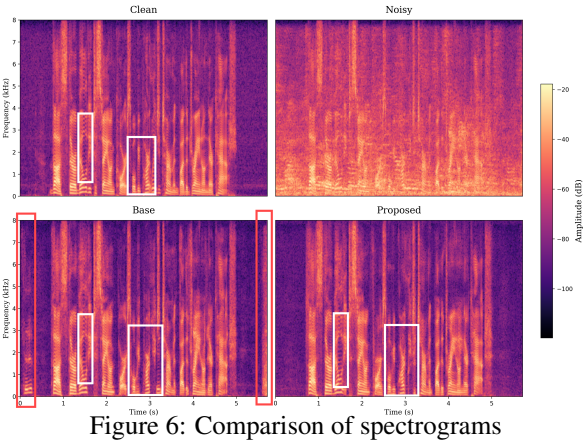

Figure 6: Comparison of spectrograms

References

[1] J. Richter *et al.*, "Speech Enhancement and Dereverberation with Diffusion-Based Generative Models," *IEEE/ACM TASLP*, vol. 31, pp. 2351–2364, Jun. 2023.

[2] J.-M. Lemercier *et al.*, "StoRM: A Diffusion-Based Stochastic Regeneration Model for Speech Enhancement and Dereverberation," *IEEE/ACM TASLP*, vol. 31, pp. 2724–2737, Jun. 2023.

[3] S. Welker *et al.*, "Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain," in *Proc. Interspeech*, 2022, pp. 2928–2932.

[4] Y.-J. Lu *et al.*, "Conditional Diffusion Probabilistic Model for Speech Enhancement," in *Proc. IEEE ICASSP*, 2022, pp. 7402–7406.

[5] H. Shi *et al.*, "Diffusion-Based Speech Enhancement with Joint Generative and Predictive Decoders," in *Proc. IEEE ICASSP*, 2024, pp. 12951–12955.

[6] H. Wang *et al.*, "Cross-Domain Diffusion Based Speech Enhancement for Very Noisy Speech," in *Proc. IEEE ICASSP*, 2023, pp. 1–5.

[7] S. Woo *et al.*, "CBAM: Convolutional Block Attention Module," in *Proc. ECCV*, 2018, pp. 3–19.

[8] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," in *Proc. ICLR*, 2021.

[9] A. Brock *et al.*, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. ICLR*, 2018.

[10] T. Karras *et al.*, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF CVPR*, 2020, pp. 8110–8119.

[11] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. NeurIPS*, 2017, pp. 5998–6008.

[12] J. S. Garofolo *et al.*, "CSR-I (WSJ0) complete," 1993.

[13] J. Barker *et al.*, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE ASRU*, 2015, pp. 504–511.

[14] A. W. Rix *et al.*, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE ICASSP*, 2001, pp. 749–752.

[15] J. Jensen *et al.*, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM TASLP*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[16] E. Vincent *et al.*, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[17] J. Le Roux *et al.*, "SDR–half-baked or well done?," in *Proc. IEEE ICASSP*, 2019, pp. 626–630.