electronics

MDPI

Article

Speech Enhancement Algorithm Based on Microphone Array and Multi-Channel Parallel GRU-CNN Network

Ji Xi [1,*], Zhe Xu [1], Weiqi Zhang [1], Yue Xie [2] and Li Zhao [3]

[1] School of Computer Information Engineering, Changzhou Institute of Technology, No. 666, Liaohe Road, Changzhou 213022, China; xuz@czust.edu.cn (Z.X.); zhangwq@czust.edu.cn (W.Z.)

[2] School of Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, China; xie-yue0109@njit.edu.cn

[3] School of Information Science and Engineering, Southeast University, Nanjing 210096, China; zhao-li@seu.edu.cn

* Correspondence: xiji@czust.edu.cn

Abstract: This paper presents an improved speech enhancement algorithm based on microphone arrays to improve speech enhancement performance in complex settings. The algorithm's model consists of two key components: the feature extraction module and the speech enhancement module. The feature extraction module processes the speech amplitude spectral features derived from STFT (short-time Fourier transform). It employs parallel GRU-CNN (Gated Recurrent Units and CNN Convolutional Neural Network) structures to capture unique channel information, and skip connections are utilized to enhance the model's convergence speed. The speech enhancement module focuses on obtaining cross-channel spatial information. By introducing an attention mechanism and applying a global hybrid pooling strategy, it reduces feature loss. This strategy dynamically assigns weights to each channel, emphasizing features that are most beneficial for speech signal restoration. Experimental results on the CHIME3 dataset show that the proposed model effectively suppresses diverse types of noise and outperforms other algorithms in improving speech quality and comprehension.

Keywords: speech enhancement; microphone array; CNN; GRU

check for updates

Academic Editor: Pawel Strumillo

Received: 3 January 2025
Revised: 1 February 2025
Accepted: 8 February 2025
Published: 10 February 2025

Citation: Xi, J.; Xu, Z.; Zhang, W.; Xie, Y.; Zhao, L. Speech Enhancement Algorithm Based on Microphone Array and Multi-Channel Parallel GRU-CNN Network. *Electronics* **2025**, *14*, 681. https://doi.org/10.3390/electronics14040681

1. Introduction

The core objective of speech enhancement technology [1] is to isolate clear and intelligible desired speech from noisy audio signals, enhance the quality of the speech signal, and establish a foundation for subsequent tasks such as speech recognition and audio-visual communication. As technology advances, the applications of speech enhancement have broadened significantly. From in-car speakerphones and video conferencing systems to hearing aids [2] and various other electronic devices, speech enhancement technology plays a crucial role.

Early limitations in technology and hardware resulted in a focus on single-channel speech enhancement methods, primarily due to their simplicity and minimal hardware requirements. Notable techniques include Wiener filtering [3], spectral subtraction [4], and Kalman filtering [5]. These methods demonstrate effective performance when the sound source is relatively stationary, making them particularly suitable for applications such as speech communication and recognition.

However, in practical environments, factors such as complex background noise, reverberation, and echo can significantly impact the effectiveness of single-channel speech enhancement techniques. Consequently, researchers have shifted their focus to microphone array speech enhancement technologies. A microphone array [6,7], consisting of multiple microphones arranged in a specific configuration, captures not only time and frequency domain information but also spatial domain information. This setup provides directionality toward the target sound source, effectively suppressing interference and ambient noise from other directions, thereby improving speech quality.

In the early stages of array signal processing, beamforming techniques were widely utilized. Beamforming leverages spatial information from multiple channels to create a spatial filter that applies varying gains to signals originating from different directions. However, practical performance can be influenced by deviations from theoretical models and localization errors. With advancements in hardware and the emergence of machine learning and deep learning theories, more sophisticated technologies have been introduced to improve the quality of output speech from microphone arrays [8].

Neural networks have significantly improved various aspects of traditional methods for enhancing speech captured by microphone arrays. A popular research direction involves utilizing deep learning techniques for parameter estimation in beamformers. For example, Luo et al. [9] introduce a time-domain beamforming model designed for low-latency scenarios, where frame-level time-domain adaptive beamforming is applied to selected reference channels, and the results are subsequently computed for all remaining channels. Sun et al. [10] employ a Convolutional Recurrent Encoder-Decoder (CRED) structure to extract spectral context and spatial information for accurate beamforming weight estimation. Additionally, neural networks [11,12] enhance conventional processes and can yield superior results across various tasks. For instance, Chau et al. [13] proposed an innovative post-filtering module that converts the preprocessed time-domain signal into an image, facilitating the extraction of richer features and thereby enhancing in-vehicle speech enhancement tasks.

Research has also focused on designing multichannel speech enhancement networks that rely solely on deep learning techniques. These algorithms enhance multi-channel speech signals directly by employing deep neural networks, either through time-domain waveform mapping [14,15] or time-frequency domain modeling [16–18]. Innovations include the integration of Spherical Harmonic Transform (SHT) features as auxiliary inputs [19], with separate decoders that extract and fuse SHT and STFT features to estimate the STFT spectrum of the desired speech signal, thereby utilizing spatial distribution information more effectively. Another approach, known as McNet [20], employs a multi-cue fusion network with modules designed to extract both full-band and narrow-band spatial information, as well as spectral information, fully integrating these features to enhance performance. Recent research has also investigated multichannel speech enhancement algorithms based on graph signal processing [13,21,22], using graph convolutional blocks to fuse relevant features for target speech estimation.

To further enhance the performance of microphone arrays in managing unknown noise situations, this paper proposes a deep convolutional neural network architecture for developing a multi-microphone speech enhancement model. This model consists of two stages: feature extraction and feature fusion. In the feature extraction stage, time and frequency domain features that reflect frequency information are extracted from the speech signal of each channel using a parallel symmetric GRU-CNN structure. In the feature fusion stage, multi-channel feature maps are integrated through hybrid pooling and channel attention weighting to more comprehensively capture the spatial domain information of the microphone array. Experimental results on the CHIME3 dataset demonstrate that this model offers superior noise reduction and improved speech quality compared to other algorithms.

2. Methodology

2.1. Model Structure

To enhance the effectiveness of speech enhancement and improve model robustness, this paper proposes an array speech enhancement algorithm that integrates Gated Recurrent Units (GRU), Convolutional Neural Networks (CNNs), and attention mechanisms. The proposed model capitalizes on the advantages of CNNs, which offer low computational overhead and comprehensive feature extraction capabilities for designing the speech enhancement system. It consists of a parallel GRU-CNN module for multi-channel feature extraction and an attention-based feature selection and fusion module. The multi-channel parallel symmetrical GRU-CNN module, based on a convolutional encoder-decoder architecture, extracts both temporal and frequency domain features from the input multi-channel speech signals. The feature selection and fusion module, utilizing an attention mechanism, performs weighted fusion of features from each channel, effectively utilizing frequency and spatial information to achieve enhanced speech quality from multiple microphones.

The framework of the established microphone array speech enhancement model is illustrated in Figure 1. First, the STFT is employed to extract the magnitude spectrum features from the M-channel microphone array signals, where F and T denote the frequency and time dimensions, respectively. Next, the speech spectral features from each microphone channel are processed through parallel GRU-CNN feature extraction modules, which capture the frequency domain feature information and contextual spectral characteristics of the input signals for each channel. During feature learning, the prediction for the current frame is generated by inputting a concatenated feature map that includes the current frame and the previous seven frames. Subsequently, the output features from the multi-channel feature extraction module are concatenated into M-channel feature maps and transmitted to the feature selection and fusion module. By utilizing hybrid pooling and attention mechanisms, the spatial information among the speech signals from each channel is utilized to dynamically assign weights to the features of each microphone channel. This process results in attention-weighted multi-channel feature maps, which are then fused, and the final predicted speech features for the current frame are produced through a convolutional layer. Finally, the speech signal is reconstructed using the Inverse Short-Time Fourier Transform (ISTFT) [23], incorporating the phase information from the reference microphone.

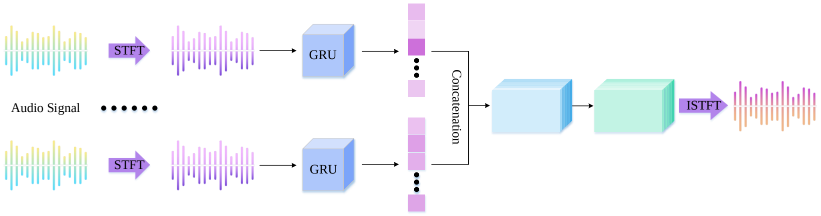

Figure 1. Multi-Channel Speech Enhancement Model Utilizing Microphone Arrays. The STFT is utilized to extract the spectral features of the speech signal [24], while the Inverse Short-Time Fourier Transform (ISTFT) is employed to reconstruct the speech signal, incorporating the phase information [23].

2.2. Amplitude Spectrum Feature Extraction

The short-time Fourier transform (STFT) [24] is employed to extract the spectral features of the speech signal, serving as the foundation for the subsequent audio noise reduction process. The feature extraction process primarily consists of two steps: applying a window to the frames and performing the STFT. Let $x_i(n)$ be the received signal from the i-th ($i = 1, 2, \ldots, N$) array element in the microphone array, and let n denote the sample number. To ensure a smooth transition of the signal, the frame length is set to L, and the frameshift is set to l. The channel speech signal $x_i(n)$ is divided into multiple single-frame signals $x_i(k \cdot l + m)$, where $0 \le k < K$, $0 \le m < L$, k is the frame number, K is the total number of decomposed frames, and m is the intra-frame sample number. By applying a Hamming window to the single-channel speech signal, spectral leakage can be further minimized. The expression for this process is as follows:

$$w(m) = \begin{cases} 0.54 - 0.46\cos[2\pi m/(L-1)], & 0 \le m < L \\ 0, & m \ge L \end{cases} \tag{1}$$

The speech signal, after windowing, can be expressed as:

$$x_i(t,m) = w(m)x_i(t \cdot l + m) \tag{2}$$

The STFT of the speech signal from this channel, after subframe and windowing, can be calculated as:

$$X_i(t,f) = \sum_{m=0}^{M-1} x_i(t,m)e^{-j\frac{2\pi mf}{M}}, \quad f = 0,1,\ldots,M-1 \tag{3}$$

$X_i(t,f)$ reflects the time-frequency characteristics of a single-channel speech signal, including both amplitude and phase information. The input feature of this paper is the amplitude spectrum, which requires a modulo operation. The speech signal, after windowing, can be expressed as:

$$M_i(t,f) = |X_i(t,f)|, \quad (i = 1,2,\ldots,N) \tag{4}$$

$M_i(t,f)$ represents the amplitude spectral signature of the i-th microphone channel based on frame-level input. The model in this paper utilizes a window length of 256 with a half-frame stack, allowing for the extraction of unique frequency information by selecting the first 129 points to minimize redundancy.

2.3. Design of Feature Calculation Module

To leverage the respective advantages of CNN and RNN, this paper employs a hybrid approach that combines CNN and GRU for the extraction of speech spectrum features. This method aims to fully capture the variations in information reflected in continuous frames of speech by utilizing the GRU's memory capabilities in the time-series dimension. Simultaneously, it capitalizes on the CNN's robust ability to extract spectral features, as CNN processes all features uniformly and extracts spatial features through convolutional kernels. The structure of the feature extraction sub-module for each channel is illustrated in Figure 2.

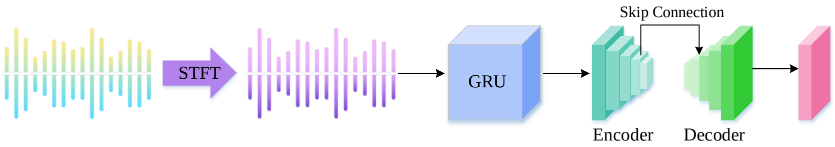

Figure 2. Structure of the GRU-CNN Feature Extraction Submodule.

The GRU-CNN feature extraction module extracts effective features from each microphone channel in parallel. Each convolutional layer is organized into a convolutional block that includes batch normalization (BN) and ReLU activation layers. The specific components of the designed feature extraction process are described as follows:

The GRU-CNN feature extraction module processes the input speech signal for each channel on a frame-by-frame basis. To fully leverage the contextual information of the input speech signal for enhanced performance, the network model does not limit itself to using only the features of the current frame. Instead, it concatenates the current frame with the features from the previous 7 frames, using this concatenated input to predict the current frame with the aid of historical information. This approach enables the model to be trained on a richer set of contextual data, thereby improving the overall efficiency of information utilization within the model.

The input to the GRU-CNN model consists of a feature map of the noisy speech spectrum, which is extracted using the STFT over 8 consecutive frames. This feature map has dimensions of (F, T, 1), where F denotes the frequency dimension and T de-notes the time dimension. The output of the feature extraction module is a feature map that captures both the frequency domain and inter-frame information of the speech signal in the current frame, with dimensions of (F, 1, N), where N denotes the number of filters in the last convolutional layer of the GRU-CNN.

The original 8 consecutive frames of noisy speech are subjected to STFT to extract spectral features. After obtaining the spectral features for each frame, they are organized into a three-dimensional spectral feature map with dimensions (F, T, 1). Since the GRU layer processes sequential data, the time series features are adapted to meet the input requirements of the GRU. The number of hidden units in the GRU is set to 64 to effectively capture long-term temporal dependencies between frames.

To maintain the spatial resolution of the feature map and preserve more details and fine-grained features of the input speech signal spectrum, the feature extraction module employs a convolutional self-encoder structure [25]. This structure consists entirely of convolutional layers and does not strictly adhere to the traditional self-encoder principle of dimensionality reduction, which typically involves "first decreasing and then in-creasing". Instead, a cyclic cascaded convolutional architecture is used to design both the encoder and decoder, resulting in a lightweight feature learning module. During the processing of multi-channel speech signals, the spectral feature maps of each channel are independently processed through the convolutional autoencoder, enabling the model to capture the unique features of each channel. This parallel convolutional self-encoder effectively learns the essential spectral feature information from each channel, providing a robust foundation for subsequent feature selection and fusion to extract spatial information.

The convolutional self-encoder within the GRU-CNN feature extraction module is structured as a fully convolutional neural network, comprising a total of 12 convolutional layers. Three of these layers are grouped, with the grouping repeated 4 times, featuring filter widths of 9, 5, and 9, and the number of filters set at 18, 30, and 8, respectively. This configuration effectively balances computational load while facilitating the extraction of intricate details. The GRU module adeptly captures contextual temporal information from the current frame of the speech signal. Unlike traditional convolutional neural networks that perform convolution across all dimensions simultaneously, this design emphasizes convolution along the frequency axis, utilizing a filter width of 1 along the time dimension for most layers. This approach ensures that important local features in the frequency dimension of the speech signal are preserved and highlighted during the forward propagation of the network. Additionally, a skip connection mechanism is incorporated in certain convolutional network architectures, such as the skip connection between the 5th and 8th convolutional layers [26]. This design provides additional pathways for the backpropagated gradients, facilitating model convergence.

2.4. Feature Selection and Fusion Module

The multi-channel parallel feature extraction module processes each channel to obtain distinct speech feature information. Since the correlation between the speech information received by each microphone and the target speech signal varies, the speech feature information from each channel does not contribute equally to the model. Microphone array speech enhancement considers not only the speech information from each channel but also incorporates the spatial information reflected by the cross-channel per-frame features. To effectively extract the cross-channel spatial information, the feature selection and fusion module designed in this paper employs a hybrid pooling and attention mechanisms for feature fusion. This approach allows for more accurate extraction of useful information for the model. The block diagram of the feature selection and fusion module is illustrated in Figure 3.

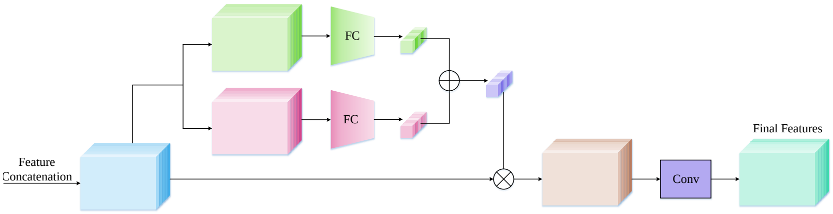

Figure 3. Speech enhancement module based on channel attention mechanism.

Inspired by the efficient channel attention mechanism of ECA-Net [27], ECA-Net employs Global Average Pooling (GAP) to downsample multi-channel feature maps and utilizes one-dimensional convolution to facilitate local cross-channel interactions among neighboring channels without reducing dimensionality. This approach yields performance improvements over traditional channel attention while adding very few parameters. However, global average pooling computes the mean value for each feature map, resulting in the loss of some feature information. To address this issue, this paper introduces a hybrid pooling method that combines Average Pooling (AP) and Maximum Pooling (MP). This hybrid approach captures the overall distribution characteristics while emphasizing significant feature information. First, each microphone channel signal $X_i \in \mathbb{R}^{F\times 1\times N}$ $(i = 1, 2, \ldots, M)$ that has undergone the feature extraction process is subjected to an inter-channel feature splicing operation to form the input feature map $X \in \mathbb{R}^{F\times M\times N}$ in the feature selection and fusion module, where F is the frequency dimension, M is the microphone channel dimension, and N is the dimension of the extracted features. Then, in order to fully utilize the spatial and frequency domain information, two different pooling strategies are applied in the feature dimension to extract the global features of the signal manifested in the frequency domain, and the squeeze operation is employed to eliminate the single-size dimension, resulting in two feature maps with dimensions (F, M). The output feature maps from the two pooling layers are dimensionally replaced and subsequently passed through a fully connected layer to yield feature maps with dimensions $(M, 1)$, which reflect the importance measures of the frequency features for each channel. The feature maps from the two branches are summed and then processed through the Softmax activation function to obtain the normalized microphone channel weights a_m $(m = 1, 2, \ldots, M)$ for the current frame. The specific process can be defined as follows:

$$s^1, s^2, \ldots, s^M = f(\mathrm{AP}(X)) + f(\mathrm{MP}(X)) \tag{5}$$

$$a_m = \mathrm{SoftMax}(s^m) = \frac{e^{s^m}}{\sum_{i=1}^{M} e^{s^i}}, \tag{6}$$

where $f(\cdot)$ denotes the fully connected layer, AP and MP denote average pooling and maximum pooling, respectively. a_m denotes the importance of each channel in the overall characterization.

The attentional mechanism flexibly adjusts the weight allocation for each channel based on the varying characteristics of the input speech signal. Through this process, critical feature mappings are assigned higher weights, while relatively less important feature mappings receive lower weights. Ultimately, these weights are applied to the corresponding channels to dynamically adjust their activation states, further refining the feature representation.The weighted multichannel speech feature maps, derived from the attention mechanism, are then connected to a convolutional layer for dimensionality reconstruction, resulting in the final predicted speech feature for the current frame, i.e., $X \in \mathbb{R}^{F \times 1}$.

2.5. Loss Function

The loss function of the proposed model comprises two distinct components: the complex compression spectral mean square error (MSE) loss and the scale-invariant signal-to-distortion ratio (SI-SDR) loss. The former can be expressed as follows:

$$L_{\text{MSE}} = \beta \text{MSE}(X^C, \hat{X}^C) + (1 - \beta)\text{MSE}(|X|^C, |\hat{X}|^C) \tag{7}$$

where $\text{MSE}(\cdot)$ denotes the MSE function, the superscript C denotes the spectrogram compression factor, and β denotes the loss weighting factor. $\hat{X}$ and X denote the enhanced speech signal and the target speech signal, respectively.

The SI-SDR loss can be expressed as follows:

$$L_{\text{SI-SDR}} = -10 \log_{10} \frac{\|X\|_2^2}{\|\alpha \hat{X} - X\|_2^2}, \tag{8}$$

where $\alpha = \frac{\|X\|_2^2}{\langle \hat{X}, X \rangle}$ is used to ensure scale invariance by normalizing $\hat{X}$ and X to zero mean before computation. The total loss for the SE model is defined as follows:

$$L_{\text{SE}} = \gamma L_{\text{MSE}} + (1 - \gamma) L_{\text{SI-SDR}} \tag{9}$$

3. Experimental Setup and Analysis

3.1. Experimental Datasets

The experiments presented in this paper utilize the widely recognized far-field multichannel dataset CHIME3. The microphone array configuration in the CHIME-3 dataset consists of six omnidirectional microphones, arranged in a typical planar array structure.

The recordings for the CHIME3 dataset were made by 12 native American English speakers aged between 20 and 50. Initially, the recordings were conducted in a near echo-free booth environment using a corpus selected from the WSJ0 corpus. Subsequently, recordings were carried out in four different noisy environments: buses (BUS), cafeterias (CAF), pedestrian streets (PED), and regular streets (STR). The CHIME-3 dataset includes both actual recorded speech and simulated speech, which was created by mixing clear speech from the WSJ0 corpus [28] with background noise from the aforementioned environments.

In this paper, experiments will be conducted using a simulated speech dataset. The training set consists of 7138 multichannel noisy speech samples, while the validation and test sets contain 1640 and 1320 multichannel simulated speech signals, respectively.

3.2. Experimental Setup

The experiments were conducted using the Python 3.8 programming language on the Windows 10 operating system. A neural network architecture was built using the TensorFlow framework, and the training process was accelerated with the NVIDIA GeForce GTX 1070Ti graphics card.

In this paper, the grid search technique is employed to select hyperparameters for the proposed model. An iterative selection of combinations within the predefined value space for each parameter is conducted to identify the parameter values that yield the best results. Ultimately, the learning rate of the proposed network model is set to 0.001. Additionally, a Dropout operation with a coefficient of 0.2 is applied between the GRU layer and the convolutional layer, as well as between the feature fusion layer and the convolutional layer. This ensures that 20% of the neurons are randomly selected to have their outputs set to zero during each training iteration, thereby excluding them from the weight update process in backpropagation. During the training phase, 100 iterations were performed with a batch size of 64. The network was optimized using the MSE loss function and the Adam [29] algorithm, with the exponential decay parameter for the first-order moment estimation β_1 set to 0.9, and the exponential decay parameter for the second-order moment estimation β_2 set to 0.999.

The performance evaluation metrics for the algorithm are the Perceptual Evaluation of Speech Quality (PESQ) [30] and the Short-Time Objective Intelligibility (STOI) [31].

In order to evaluate the performance of the proposed model, assess the effectiveness of its constituent modules, compare the experimental results with those of other state-of-the-art algorithms, and to clarify the speech enhancement effects of the model, this paper conducts the following three experiments:

Experiment 1: The designed model is trained and tested on the CHiME3 dataset, and the performance of the algorithm is compared and analyzed with that of the traditional beamforming algorithm.

Experiment 2: Comparison experiments are conducted on the same dataset by either retaining or removing the GRU module to verify its validity within the model. Additionally, experiments are performed by selecting different feature fusion module structures to evaluate the impact of these structures on the speech enhancement performance.

Experiment 3: The proposed model is compared and analyzed against other deep learning algorithms using the same dataset. The effectiveness of these three experimental tasks will be evaluated based on two objective evaluation metrics: PESQ and STOI.

3.3. Performance Comparison with Competing Algorithms

The experiments were initially conducted using the CHIME3 dataset to evaluate various beamforming algorithms, with the results presented in Figure 4. The algorithms compared include the traditional generalized collateral phase cancellation (GSC), the GSC algorithm based on joint speech feature adaptive control (GSC-SPP), the GSC with a cascaded posterior GRU network (GSC-GRU) and the proposed model. As illustrated in the figure, the proposed model outperforms both the traditional GSC algorithm and the enhanced GSC algorithm, demonstrating improvements in the PESQ and STOI metrics across all noisy environmental scenarios. This indicates a significant enhancement in both speech quality and intelligibility. The primary reason for this improvement is that the proposed model effectively reconstructs the spectral details of speech in noisy environments and excels at suppressing residual noise. The proposed model successfully eliminates noise signals in the low-frequency band while a providing superior suppression of high-frequency noise in the non-speech segments, resulting in enhanced speech quality and improved intelligibility.

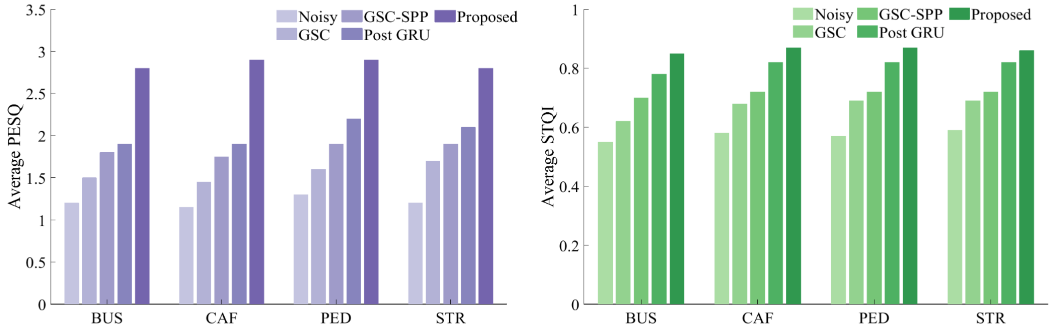

Figure 4. Comparison of algorithm processing effect under different types of noise.

3.4. Ablation Study

3.4.1. Impact of the GRU Module on Model Performance

To verify the impact of the GRU module on temporal feature extraction, this paper presents a validation experiment comparing the performance of the algorithm with and without the GRU module. The experimental results are summarized in Table 1, where CNN_NoGRU denotes the network model with the GRU network removed. The table indicates that the proposed model incorporating the GRU module improves the PESQ and STOI scores by 0.6278 and 0.0352, respectively. When the front GRU network is omitted, the results of the two objective evaluation metrics decline. This decrease is primarily due to the GRU's effectiveness in capturing long-term dependencies in time series data, which is particularly crucial for extracting features from speech signals, as they often contain significant information that spans multiple time steps. Since the feature extraction module in this paper processes data on a frame-by-frame basis, the convolutional layer in the convolutional self-encoder is specifically designed to capture frequency domain feature information solely along the frequency axis. The incorporation of a memory function for time-series features, following the addition of the GRU, enhances the comprehensiveness of the features extracted by the model's feature ex-traction module. Additionally, the design of the GRU network remains relatively straightforward, resulting in a smaller number of parameters.

Table 1. Comparison of PESQ and STOI with and without gated cyclic cell networks.

Model	PESQ	STOI
CNN_NoGRU	2.2378	0.7493
Proposed	2.8656	0.7845

3.4.2. Impact of Speech Enhancement Module on Model Performance

To verify the contributions of hybrid pooling and attentional weighting to the model's effectiveness, this section discusses the feature selection fusion module in a categorized manner. This includes the convolutional self-coder (GRU-CNN) without the speech enhancement module, as well as models that incorporate simple speech enhancement modules: Maximum Pooling (GRU-CNN-MP) and Average Pooling (GRU-CNN-AP). Additionally, the model that integrates the Attention Mechanism feature selection fusion module is referred to as the Proposed model. Each model with different feature selection fusion modules was trained and tested, and Figure 5 presents the results of the objective evaluation metrics set for the various speech enhancement modules.

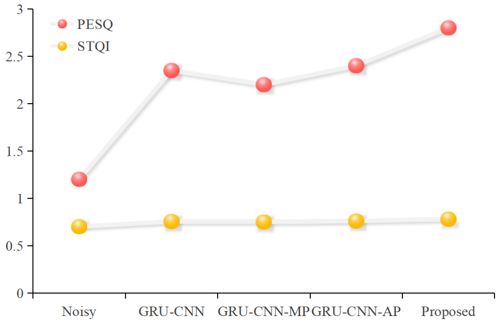

Figure 5. Performance comparison of different speech enhancement modules.

The experimental results indicate a significant improvement in both speech quality and intelligibility due to the speech enhancement module in the proposed model. Specifically, the PESQ score increased by 0.5813, and the STOI score is improved by 0.05 compared to the model without the speech enhancement module. This enhancement can be attributed to two main factors: (1) global features are obtained by hybrid pooling, which improves the index results compared with that of the single-pooling approach. The reason is that single pooling is too absolute for the feature selection of the feature channel and does not fully consider the balance of average and significant features; (2) adding the attention mechanism can dynamically assign weights to the spatial dimension of the convolutional output feature map, and focus on the feature fusion according to the attention weights, which improves the ability of the neural network for spatial feature information. Therefore, adding the hybrid pooling module and attention mechanism to the network model can further acquire cross-channel spatial information and improve the multichannel speech enhancement performance.

3.5. Performance Comparison with Neural Network Algorithms

In order to further validate the effect of the proposed model presented in this paper, this section compares the proposed model with other deep learning methods [32,33]. The different deep learning models were trained and tested on the CHIME3 multichannel speech dataset. The experimental results are shown in Table 2.

Table 2. Performance comparison of different deep learning algorithms, where the bold number indicates the best score.

Model	PESQ	STOI	GFLOPs	Parameter Quantity (M)
Noisy	1.2133	0.6585	12.28	46.42
CRN [32]	2.1823	0.7645	12.48	48.24
CAU-Net [34]	2.3956	0.7722	13.73	46.28
FT-JNF [35]	2.5656	0.7755	12.16	44.85
McNet [20]	2.6856	0.7845	11.85	44.18
SpatialNet [36]	2.8356	**0.7915**	14.55	52.22
DeFT-AN [37]	2.8256	0.7805	11.55	44.36
Proposed	**2.8656**	0.7845	**11.30**	**43.85**

As shown in Table 2, within the same dataset, compared with the noisy speech signal, the PESQ and STOI of the proposed model are improved by 1.6523 and 0.126, respectively. The PESQ and STOI of the processed speech are improved by 0.03 than that of the SpatialNet model, and the STOI of the proposed model is slightly less than that of the densely-connected SpatialNet model, and the perceived quality of the speech is slightly insufficient, the reason is analyzed that the proposed model has better suppression effect on noise, which makes the speech suffer some loss in some frames where speech and noise coexist. However, the SpatialNet model based on the channel attention mechanism has a large number of dense connections, and during the training and calculation process, as the network layer deepens, the dimension of the channel will also be expanded, which will result in a relatively large amount of computation, while the proposed model has a good performance in terms of performance and computation.

In addition, as shown in Table 2, the proposed method demonstrates a reduction in both computational cost and the number of parameters. Compared to other methods, the approach presented in this study achieves a more favorable balance between performance and computational cost. These superior results can be attributed to the integration of multiple pooling strategies that effectively extract key features, as well as the implementation of an attention mechanism that captures global contextual information. This enables the model to achieve better outcomes without requiring extensive computational resources.

4. Conclusions

In this paper, we propose an improved multichannel speech enhancement model based on a deep learning network. The model incorporates a convolutional self-encoder and a feature selection and fusion module, creatively embedding hybrid pooling and attention mechanisms to fully utilize the detailed frequency-domain characteristics of the microphone array speech signals and achieve better noise reduction. Experiments conducted on the CHIME3 dataset demonstrate that the algorithm excels in leveraging the unique frequency information of each channel and the related spatial domain in-formation between channels, outperforming traditional algorithms in mitigating the impact of non-smooth and non-coherent noise on speech signals. Compared to conventional methods, our algorithm significantly improves speech quality and intelligibility, offering a more robust array speech enhancement effect.

Author Contributions: Conceptualization, J.X. and Z.X.; methodology, J.X. and W.Z.; coding, W.Z.; validation, J.X.; investigation, Y.X.; writing—original draft preparation, J.X. and Z.X.; writing—review and editing, J.X. and Z.X.; visualization, W.Z.; supervision, J.X.; project administration, L.Z.; funding acquisition Z.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Science and Technology Plan Project of Changzhou (CJ20220151), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (23KJA520001).

Data Availability Statement: The CHIME3 dataset is available at https://www.chimechallenge.org/challenges/chime3/index (accessed on 17 December 2015).

Acknowledgments: Thanks to the anonymous reviewers and editor from the journal.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Loizou, P. *Speech Enhancement: Theory and Practice*; CRC Press: Boca Raton, FL, USA, 2007.
2. Chhetri, S.; Joshi, M.S.; Mahamuni, C.V.; Sangeetha, R.N.; Roy, T. Speech Enhancement: A Survey of Approaches and Applications. In Proceedings of the 2023 2nd International Conference on Edge Computing and Applications (ICECAA), Namakkal, India, 19–21 July 2023; pp. 848–856.
3. Lim, J.S.; Oppenheim, A.V. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE* **1979**, *67*, 1586–1604. [CrossRef]
4. Boll, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, *27*, 113–120. [CrossRef]
5. Paliwal, K.; Basu, A. A speech enhancement method based on Kalman filtering. In Proceedings of the ICASSP'87, IEEE International Conference on Acoustics, Speech, and Signal Processing, Dallas, TX, USA, 6–9 April 1987; Volume 12, pp. 177–180.

6. Marques, I.; Sousa, J.; Sá, B.; Costa, D.; Sousa, P.; Pereira, S.; Santos, A.; Lima, C.; Hammerschmidt, N.; Pinto, S.; et al. Microphone array for speaker localization and identification in shared autonomous vehicles. *Electronics* **2022**, *11*, 766. [CrossRef]
7. Grumiaux, P.A.; Kitić, S.; Girin, L.; Guérin, A. A survey of sound source localization with deep learning methods. *J. Acoust. Soc. Am.* **2022**, *152*, 107–151. [CrossRef]
8. Mehrish, A.; Majumder, N.; Bharadwaj, R.; Mihalcea, R.; Poria, S. A review of deep learning techniques for speech processing. *Inf. Fusion* **2023**, *99*, 101869. [CrossRef]
9. Luo, Y.; Han, C.; Mesgarani, N.; Ceolini, E.; Liu, S.C. FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 260–267.
10. Sun, S.; Jin, J.; Han, Z.; Xia, X.; Chen, L.; Xiao, Y.; Ding, P.; Song, S.; Togneri, R.; Zhang, H. A Lightweight Fourier Convolutional Attention Encoder for Multi-Channel Speech Enhancement. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
11. Li, G.; Liang, S.; Nie, S.; Liu, W.; Yang, Z.; Xiao, L. Deep Neural Network-Based Generalized Sidelobe Canceller for Robust Multi-Channel Speech Recognition. In Proceedings of the INTERSPEECH, Shanghai, China, 25–29 October 2020; pp. 51–55.
12. Zhou, Y.; Bao, C.; Cheng, R. GSC based speech enhancement with generative adversarial network. In Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 901–906.
13. Chau, H.N.; Bui, T.D.; Nguyen, H.B.; Duong, T.T.H.; Nguyen, Q.C. A novel approach to multi-channel speech enhancement based on graph neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 1133–1144. [CrossRef]
14. Luo, Y.; Chen, Z.; Mesgarani, N.; Yoshioka, T. End-to-end microphone permutation and number invariant multi-channel speech separation. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6394–6398.
15. Liu, C.L.; Fu, S.W.; Li, Y.J.; Huang, J.W.; Wang, H.M.; Tsao, Y. Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1888–1900. [CrossRef]
16. Gu, R.; Zhang, S.X.; Chen, L.; Xu, Y.; Yu, M.; Su, D.; Zou, Y.; Yu, D. Enhancing end-to-end multi-channel speech separation via spatial feature learning. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7319–7323.
17. Wang, Z.Q.; Wang, P.; Wang, D. Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 2001–2014. [CrossRef]
18. Tan, K.; Wang, Z.Q.; Wang, D. Neural spectrospatial filtering. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 605–621. [CrossRef]
19. Pan, J.; Shen, P.; Zhang, H.; Zhang, X. Efficient Multi-Channel Speech Enhancement with Spherical Harmonics Injection for Directional Encoding. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 9561–9565.
20. Yang, Y.; Quan, C.; Li, X. McNet: Fuse multiple cues for multichannel speech enhancement. In Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
21. Tzirakis, P.; Kumar, A.; Donley, J. Multi-channel speech enhancement using graph neural networks. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 3415–3419.
22. Hao, M.; Yu, J.; Zhang, L. Spatial-temporal graph convolution network for multichannel speech enhancement. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 7–13 May 2022; pp. 6512–6516.
23. Yang, B. A study of inverse short-time Fourier transform. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 3541–3544.
24. Benesty, J.; Chen, J.; Habets, E.A. *Speech Enhancement in the STFT Domain*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2011.
25. Masci, J.; Meier, U.; Cireşan, D.; Schmidhuber, J. Stacked convolutional auto-encoders for hierarchical feature extraction. In *Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, 14–17 June 2011, Proceedings, Part I 21*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 52–59.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
27. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.

28. Paul, D.B.; Baker, J. The design for the Wall Street Journal-based CSR corpus. In Proceedings of the Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, NY, USA, 23–26 February 1992.
29. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
30. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), Salt Lake City, UT, USA, 7–11 May 2001; Volume 2, pp. 749–752.
31. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 15–19 March 2010; pp. 4214–4217.
32. Tan, K.; Wang, D. A convolutional recurrent neural network for real-time speech enhancement. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; Volume 2018, pp. 3229–3233.
33. Barker, J.; Marxer, R.; Vincent, E.; Watanabe, S. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, USA, 13–17 December 2015; pp. 504–511.
34. Tolooshams, B.; Giri, R.; Song, A.H.; Isik, U.; Krishnaswamy, A. Channel-attention dense u-net for multichannel speech enhancement. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 836–840.
35. Tesch, K.; Gerkmann, T. Insights into deep non-linear filters for improved multi-channel speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *31*, 563–575. [CrossRef]
36. Quan, C.; Li, X. SpatialNet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 1310–1323. [CrossRef]
37. Lee, D.; Choi, J.W. DeFT-AN: Dense frequency-time attentive network for multichannel speech enhancement. *IEEE Signal Process. Lett.* **2023**, *30*, 155–159. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.