Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 9453911, 9 pages
https://doi.org/10.1155/2021/9453911

Research Article

Multichannel Speech Enhancement in Vehicle Environment Based on Interchannel Attention Mechanism

Xueli Shen,[1,2] **Zhenxing Liang**,[2] **Shiyin Li**,[1] **and Yanji Jiang**[2,3]

[1]*School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China*
[2]*School of Software, Liaoning Technical University, Huludao 125105, China*
[3]*Suzhou Automotive Research Institute, Tsinghua University, Suzhou 215100, China*

Correspondence should be addressed to Yanji Jiang; jyjvip@126.com

Received 6 August 2021; Revised 27 September 2021; Accepted 8 October 2021; Published 15 November 2021

Academic Editor: Geqi Qi

Speech enhancement in a vehicle environment remains a challenging task for the complex noise. The paper presents a feature extraction method that we use interchannel attention mechanism frame by frame for learning spatial features directly from the multichannel speech waveforms. The spatial features of the individual signals learned through the proposed method are provided as an input so that the two-stage BiLSTM network is trained to perform adaptive spatial filtering as time-domain filters spanning signal channels. The two-stage BiLSTM network is capable of local and global features extracting and reaches competitive results. Using scenarios and data based on car cockpit simulations, in contrast to other methods that extract the feature from multichannel data, the results show the proposed method has a significant performance in terms of all SDR, SI-SNR, PESQ, and STOI.

1. Introduction

In the process of driving, the speech signals recorded by a microphone are often corrupted by reverberation and background noise, such as wind noise, engine noise, and tire noise, leading to considerable degradation in speech quality, particularly at low signal-to-noise ratios (SNRs) [1]. Speech enhancement technology can improve the speech quality of the interphone system and the ability of the speech recognition system. Multichannel enhancement in vehicle scenarios uses microphone arrays that are convenient and flexible for speech-enabled applications [2]. The multichannel structure could provide more spatial information from the interchannel data and better results than the signal channel.

Although the technology of microphone array has been developed for a long time, multichannel speech enhancement is still a great challenge in the field of speech recognition. The methods can be divided into two categories: one is based on the frequency domain, and the other is based on the time domain. Researchers mostly use the frequency-domain methods, which are based on the short-time spectrum estimation. Chakrabarty and Habets [3] proposed a multichannel online speech enhancement method based on time-frequency masking. Convolutional recurrent neural network (CRNN) is used to estimate the mask, and the effects of the ideal ratio mask (IRM) and ideal binary mask (IBM) on the results are discussed. The results show that the method is robust to different angles of sound sources. In [4], a multichannel speech enhancement system based on a deep neural network is proposed. Firstly, the audio signal is transformed into the frequency domain by STFT, the time-frequency mask is estimated by DNN, and the multichannel Wiener filtering is performed by using the power spectral density of speech and noise. The experimental results show that the method is effective. A beamforming method different from the traditional DNN is proposed in [5]. The spectrum of each channel is mapped to the non-Euclidean space, usually using the phase information to improve real-time performance, and the graph neural network is used for end-to-end training. Compared with the existing methods, the experiment result is better. A time-domain beamforming method named FaSNet (Filter and Sum Network) suitable for the low delay is proposed in [6]. The author selects the reference channel for filtering, calculates the filter of other channels by the reference channel, and then adds the filtered

speech of each channel as the denoised speech. The model size of the algorithm is small, and the performance is better than that of traditional beamforming methods. In [7], a streaming speech enhancement system is proposed, which adopts the Wave-U-Net framework, adds temporal convolution and attention mechanism into the encoding and decoding structure, and explores the history caching mechanism. This method achieves almost the same noise reduction effect as the nonstreaming model. The time-domain convolutional denoising autoencoders (TCDAEs) method is proposed in [8]. It is used to learn the mapping structure between noisy speech waveform and clean speech waveform and solve the problem of speech signal delay between different channels effectively. Compared with the traditional denoising autoencoder, the effect has been significantly improved.

The multichannel speech enhancement model has the most significant advantage of obtaining abundant information between channels compared with the single channel. Therefore, for the multichannel model, the way that extracts the spatial features between channels more effectively becomes the key to achieving better performance. In [9], a multichannel convolution sum (MCS) is used to extract features between channels. On the contrary, in [9], inspired by the IPD [10] feature, the interchannel convolution feature (ICD) is proposed. The method is to perform one-dimensional convolution subtraction on a pair of microphones. Based on GCC-PHAT, [6, 11] considered the normalized cross-correlation (NCC) method, which uses cosine similarity to calculate the information between channels. All the above methods achieve better performance improvement for multichannel speech enhancement. To address the problem of speech enhancement in the car cockpit, this paper proposes a novel method based on interchannel attention mechanism frame by frame (IAF), which helps analyse the influence of each channel on speech signal by using the characteristic information of the channel. Moreover, the proposed method also explores interchannel relationships and achieves more information representation on channel structure. It provides a new idea for multichannel speech enhancement based on vehicle environment and can also be applied to smart homes, teleconference, and other scenes.

The main contents of this paper are as follows: Section 1 introduces the related research work in this field. The structure of the multichannel speech enhancement model based on IAF is proposed in Section 2. The algorithm performance in the vehicle environment is evaluated in Section 3. The experimental results of the algorithm on several microphone arrays are analysed and discussed in Section 4, and Section 5 draws the conclusion and points out the future of the research work.

2. Problem Formulation

The proposed method aims to obtain an accurate estimate of the features for all the channels of a single time frame, given the input feature representation of the corresponding frame. The multichannel speech enhancement process of vehicle data is divided into four successive steps. First, spatial features from multichannel data added context information is extracted by IAF. Then, the frame-level beamforming filters are estimated by a well-trained two-stage BiLSTM model using spatial features, and the original waveforms computed by 1-dimensional convolution, for $N\,(N>=2)$ microphones, N beamforming filters are estimated. Next, the filters are adopted to filter the noisy speech in every channel, thereby obtaining the beamformed speech. Finally, add the beamformed speech as the denoised speech. The detail is presented in the following sections. A block diagram of the proposed multichannel enhancement framework is shown in Figure 1.

2.1. Data Preprocessing. It is assumed that the input signal corresponding to each microphone is represented as (2). Here, the frame length is M, the frameshift is $K \in [0, M-1]$, the total length of the speech signal is l, and the total number of frames is Z:

$$Z = \frac{l}{K} + 2, \tag{1}$$

$$x_t^i = x^i[tK: tK + M - 1],\ t \in [0, Z], \quad i = 1, \ldots, N, \tag{2}$$

where t is the frame index value and i is the channel index. $x_t^i \in \mathbb{R}^{1\times M}$ indicates that the signal vector of frame t is collected by microphone i.

Due to the different distance between each microphone and the sound source, there is a time delay between the signals received by each microphone. Add a context window to make sure the model can capture interchannel delays of signal samples [12]. We add a group of contextual speech information in x_t^i and define it as $\overline{x}_t^i$:

$$\overline{x}_t^i = x^i[tK - W: tK + W + M - 1], \tag{3}$$

where W is the size of the context window and $\overline{x}_t^i \in \mathbb{R}^{2W+M}$ is the signal vector of the microphone i containing the context information at frame t. The input sequence to these networks consists of W past and W future time frame.

2.2. Interchannel Attention Mechanism Frame by Frame. We calculate the corresponding weights of different parts of the channel by constructing the score function to describe the transmission characteristics of the signal in the channel. The principle of interchannel attention mechanism frame by frame is shown in Figure 2.

In order to extend the context information $\overline{x}_t^i$, firstly, average pooling is performed in the frame length dimension:

$$F_a = z_t^i = \frac{1}{2W + M} \sum_{j=tK-W}^{tK+W+M-1} \overline{x}^i[j], \quad i = 1, \ldots, N. \tag{4}$$

$z_t^i \in \mathbb{R}^1$ is the average value of microphone i at the number of frames t. Then, the results are input into multiple fully connected layers:

$$F_b = G_t = S\Big(P\Big(\Big[z_t^1, z_t^2, \ldots, z_t^N\Big]\Big)\Big), \quad G_t \in [0, 1]. \tag{5}$$

$G_t \in \mathbb{R}^{1\times N}$ is the microphone array feature at the frame t. $P(*)$ is a set of fully connected layers with parameter

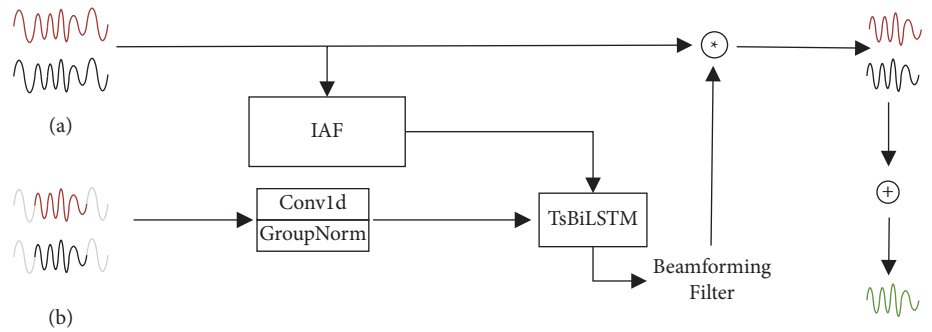


Figure 1: The overall process of multichannel speech enhancement model based on IAF. (a) The speech signal with context. (b) The original speech signal. "IAF" and "TsBiLSTM" denote the feature extraction and feature filter using the TsBiLSTM model, respectively.

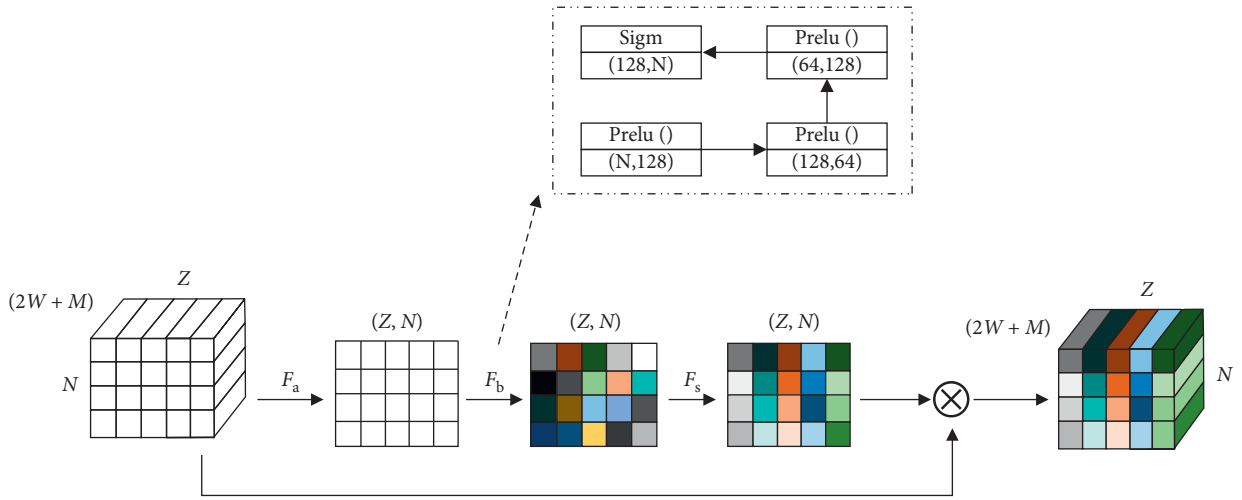


Figure 2: The module of the frame-level interchannel attention mechanism: different colors represent different weight values, and multiply the weight values with the original data.

modified linear unit (PReLU) activation function, $S(*)$ is a set of fully connected layers with sigmoid activation function, and the output of $P(*)$ and $S(*)$ are [128, 64, 128] and [N].\, respectively. Then, input G_t into the softmax activation function:

$$F_s = \overline{F}_t = \text{Softmax}(G_t), \tag{6}$$

where $\overline{F}_t$ is a vector whose sum is 1. The final output out is obtained by multiplying with $\overline{F}_t$ and x_t:

$$\text{out}_t^i = \overline{F}_t^i \times x_t^i, \quad i = 1, \ldots, N. \tag{7}$$

$\text{out}_t^i \in \mathbb{R}^{1\times(2W+M)}$ presents the speech feature sequence of the t -th frame data in the i -th channel.

By using the attention mechanism frame by frame of the speech signal in multiple channels, the model could learn the characteristics of each channel and capture spatial features between channels more accurately.

2.3. Two-Stage BiLSTM Network. The two-stage bidirectional LSTM (TsBiLSTM) is used to derive a beamformer as BiLSTM is adopted to estimate the global feature. For the beamformer, the approach aims to improve the SNR without destroying the target speech.

Figure 3 shows the TsBiLSTM architecture employed in this work. We divide the data into blocks, consider using the BiLSTM network model to obtain local and global features of the blocks and establish the timing relationship of the signal, and use the residual connections to alleviate the gradient dispersion problem.

In this work, we combine the speech signal with context information in the first stage. The observed signal can be expressed as follows:

$$y_t = \text{GroupNorm}(\text{Conv1 } d(x_t)), \tag{8}$$

$$xb_t = \text{concat}([\text{out}_t, y_t]). \tag{9}$$

$xb_t \in \mathbb{R}^{N\times 2(M+W)}$ represents the speech features of frame t and $xb \in \mathbb{R}^{Z\times N\times 2(M+W)}$ represents all the speech features; then, we do the one-dimensional convolution on xb:

$$c = \text{Conv1 } d(xb), \tag{10}$$

where $c \in \mathbb{R}^{Z\times(N\times M)}$, then divide c into S blocks of the same size. Each block presents $B_s \in \mathbb{R}^{U\times(N\times M)}, s \in [1, S]$, and all the blocks will be connected to form a four-dimensional vector $O \in \mathbb{R}^{N\times M\times S\times U}$.

We transform the shape $O \in \mathbb{R}^{N\times M\times S\times U}$ to $O \in \mathbb{R}^{(S\times N)\times U}$ $\times M$ and input the first BiLSTM:

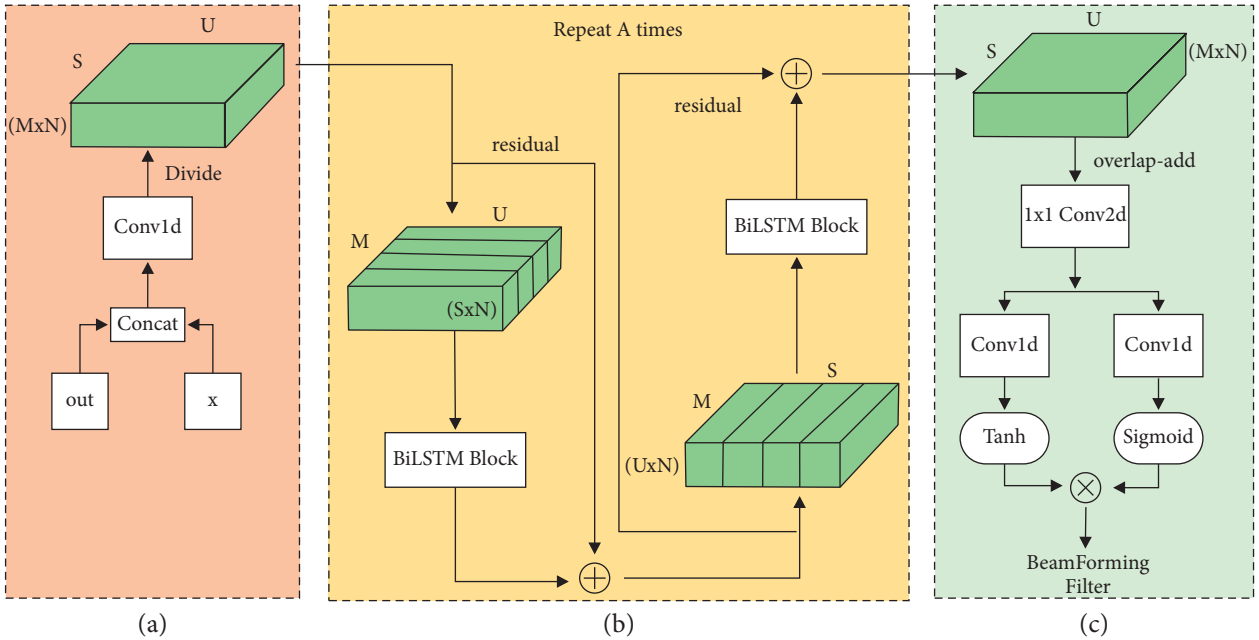


FIGURE 3: The structure of the TsBiLSTM module and illustration of the proposed module architecture. (a) The input of the TsBiLSTM includes interchannel features and original waveform. (b) The processing chain shows the two-stage BiLSTM with residual connections. (c) Carry out 1×1 convolution on the output after overlap-add method operation, then operate conv1d layer with sigmoid and tanh activation function, respectively, and obtain beamforming filter.

$$\begin{aligned} \text{out} &= \text{reshape}(O), \\ \text{out}_1 &= \text{GroupNorm}(\text{Linear}(\text{BiLSTM}(\text{out}))), \\ \text{out} &= \text{reshape}(out_1) + O. \end{aligned} \tag{11}$$

The output of BiLSTM passes through the linear layer and GroupNorm operation and then output $\text{out}_1 \in \mathbb{R}^{(S\times N)\times U\times M}$. Reshape $\text{out}_1 \in \mathbb{R}^{(S\times N)\times U\times M}$ to $\text{out}_1 \in \mathbb{R}^{N\times M\times S\times U}$, add the vector using the residual connection to reduce the problem of gradient disappearance or gradient explosion, and finally obtain $\text{out} \in \mathbb{R}^{N\times M\times S\times U}$.

In the next stage, change $\text{out} \in \mathbb{R}^{N\times M\times S\times U}$ into $\text{out} \in \mathbb{R}^{(U\times N)\times S\times M}$, then input the next BiLSTM, as the first BiLSTM block, and finally, obtain $\text{out} \in \mathbb{R}^{N\times M\times S\times U}$. Because the signals are transmitted to the BiLSTM model in different block forms, we can obtain the local and global features of the signals, respectively:

$$\begin{aligned} \text{out}_1 &= \text{reshape}(\text{out}), \\ \text{out}_2 &= \text{GroupNorm}(\text{Linear}(\text{BiLSTM}(\text{out}_1))), \\ \text{out} &= \text{reshape}(out_2) + \text{out}. \end{aligned} \tag{12}$$

Then, use the overlap-add operation to convert the segmented block back to the original sequence:

$$\text{out}_3 = O\ D(\text{out}), \tag{13}$$

where $\text{out}_3 \in \mathbb{R}^{Z\times(N\times M)}$, $O\ D(\cdot)$ is the overlap-add method, which means to restore the partitioned data. Then, convolve $\text{out}_3 \in \mathbb{R}^{Z\times(N\times M)}$ in two dimensions with the convolution kernel of size set one:

$$\text{out}_4 = \text{Conv2}\ d(\text{out}_3), \tag{14}$$

where $\text{out}_4 \in \mathbb{R}^{Z\times(N\times M)}$. We perform twice one-dimensional convolution operations on out_4, then use the activation function of Tanh and Sigmoid, respectively, and multiply the results to get the filters for each channel:

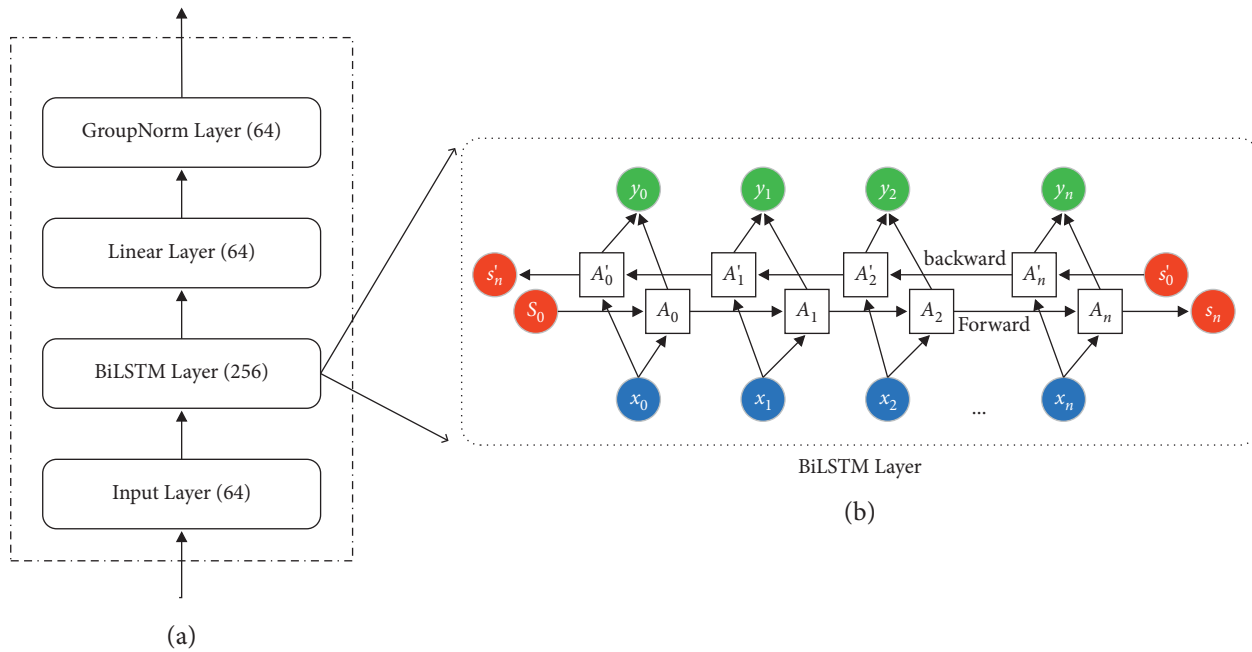
$$h = \text{Tanh}(\text{Conv1}\ d(\text{out}_4)) \odot \text{Sigmoid}(\text{Conv1}\ d(\text{out}_4)), \tag{15}$$

where $h \in \mathbb{R}^{N\times Z\times(2W+1)}$, $\odot$ is the Hadamard product symbol, $\text{Tanh}(\cdot) \odot \text{Sigmoid}(\cdot)$ is the gating mechanism of filter that controls the output data.

Figure 4 shows the structure of the BiLSTM block. The input layer is the feature vector of noisy speech with dimension 64, which is input into the BiLSTM layer with dimension 128. The output dimension is 256 since bidirectional LSTM is used. Then, input the linear hidden layer of 64, and get the output after the GroupNorm operation.

2.4. Summation. Integrating the signals of multiple channels into one signal output is an important step in the multichannel speech enhancement problem. After passing the signals of each channel through the channel filter, the results obtained are summed and averaged, that is, the final enhanced speech signal:

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} h^i \circledast \overline{x}^i, \quad i = 1, \ldots, N, \tag{16}$$



(a) (b)

FIGURE 4: The structure of the BiLSTM block. (a) The module contains BiLSTM, linear, and GroupNorm layers, where the numbers in parentheses indicate the size of the output dimension. (b) The internal structure diagram of the BiLSTM layer, where A refers to the LSTM module, x is the input data, s is the output of the hidden layer, and y is the output result BiLSTM.

where $\overline{y} \in \mathbb{R}^{N\times Z\times M}$ and $\circledast$ is convolution operation. Ultimately, $\overline{y}$ is inversed from segmentation into an enhanced speech waveform by overlapping.

2.5. Loss Function. In training and evaluation, the scale-invariant source-to-noise ratio (SI-SNR) is used as the loss function.

$$\begin{aligned} s_{\text{target}} &= \frac{\langle \overline{x}, x\rangle x}{\|x\|^2}, \\ e_{\text{noise}} &= \overline{x} - s_{\text{target}}, \\ SI-\text{SNR} &= 10\log_{10}\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2}, \end{aligned} \tag{17}$$

where $\overline{x}$ is the denoised speech and x is pure speech signal.

3. Experiment Section

The speech enhancement tasks are evaluated in four kinds of microphone array structures to simulate the location of the microphone in the car. The speech source and locations of the noise source are shown in Figure 5, where the black circle represents the microphones, the green square represents the speech source, and the red five-pointed star represents the noise source. The design of the microphone array is as follows:

(i) Consider a uniform linear array with 2 microphones with intermicrophone distance of 3 cm, and the microphone array is located in the front of the car cockpit, as Figure 5(a)

(ii) Consider a uniform linear array with 2-uniform linear distributed 2-channel microphone array with intermicrophone distance of 3 cm, and the microphone array is located in the front and middle of the car cockpit, respectively, as shown in Figure 5(b)

(iii) Consider a uniform linear array with 4 microphones with intermicrophone distance of 3 cm, and the microphone array is located in the front of the car cockpit, as shown in Figure 5(c)

(iv) Consider a distributed array with 4 microphones with intermicrophone distance of 80 cm, and the microphone is located around the car cockpit, as shown in Figure 5(d)

Different microphone array structures can reflect different spatial characteristics. In order to make the method independent of the spatial position of the required speech source, each microphone array position and source-array distance are considered under the training condition.

3.1. Datasets Building. For training, we used 3000 randomly chosen speech utterances from the LibriSpeech [13] dataset which are open and well-studied dataset used for speech enhancement, each 4 s long, with sampling frequency of 16 k Hz, and 500 were used as a validation set. Volvo car noise [14] was added to the training data as noisy speech in the car cockpit with randomly chosen SNRs between −10 dB and −5 dB. Additionally, since the number of noise is small, spsquare noise [15] as a noise source, with randomly chosen SNRs between −10 dB and −5 dB, was also added. All dataset are divided into frames with 64 sampling points length, 50% overlapping, and the context window is 256.

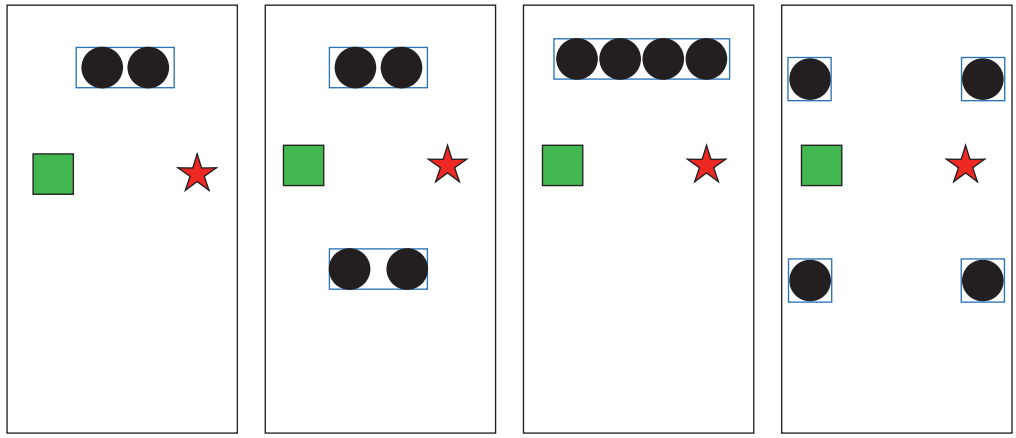


Figure 5: The distribution map of four microphone arrays. (a) Uniform linear 2-channel microphone array. (b) Two-uniform linear distributed 2-channel microphone array. (c) Uniform linear 4-channel microphone array. (d) Distributed 4-channel microphone array.

To simulate a car cockpit, we designed the space size to be 3.4 meters long, 1.8 meters wide, and 1.4 meters high. The vehicle cockpits impulse responses required to simulate real acoustic conditions are generated by gpuRIR toolbox [16], with the reverberation time (T60) selected from 0.1 seconds to 0.3 seconds randomly.

3.2. Experiment Settings. The experiment aims to verify the generalization capability of the proposed method over different microphone arrays and compare the performance to that of traditional beamformers. For a fair comparison, we make the comparison as all models, including NCC, MCS, and ICD, are based on the same two-stage BiLSTM modules presented in Section 2.3 for each microphone channel. The architecture of each BiLSTM network consists of 128 hidden layers. Set layer number 4. For MCS and ICD, the size of the convolution kernel is 64, the number of convolution kernels is 16, the step size is 2, and the expansion number is 2, leading to the output where the filter estimate for each microphone is obtained.

The BiLSTM network was trained using the Adam-based optimizer, with minibatches of 128 input signals and a learning rate of 0.001. Meanwhile, the L2 norm of 5 is used for gradient pruning to prevent gradient explosion. During training, if the loss value of the latest 10 epoch model does not decrease on the validation sets, the training will stop automatically. Dynamic strategy warmup [17] is used to adjust the learning rate during the training. This operation can warm up the model at a small learning rate in the initial stage to increase the stability of the model and then gradually reduce it with a decay rate of 0.98 every 2 epochs. The specific approach is similar to [18]. All the implementations were done in PyTorch:

$$\begin{aligned} lr &= a_1 \cdot n \cdot d_{\text{model}}^{-0.5} \cdot n\_\text{warmups}^{-1.5}, n \le n\_\text{warmups}, \\ lr &= a_2 \cdot 0.98^{[\text{epoch}/2]}, n > n\_\text{warmups}, \end{aligned} \tag{18}$$

where n is the number of training steps anda_1, a_2, warmups, and d_{model} present the hyperparameter. In the experiment, we set $a_1 = 0.2$, $a_2 = 1e^{-3}$, n_warmups = 4000 , and $d_{\text{model}} = 64$.

4. Results and Discussion

Following the common speech enhancement metrics, we adopt average SI-SNR, SDR, PESQ, and STOI improvement to evaluate the performance of multichannel speech enhancement. For a more comprehensive evaluation of the speech quality, we also report the performances under different SNRs of speech and noise to give a more comprehensive model assessment. The experimental results are summarized in Table 1, where the highlighted numbers with black are the best scores for each model. The results indicate that the performance of proposed method is better than other methods when tested at different SNRs, which verifies the effectiveness of the model. By assigning weight values to each channel frame by frame, using attention mechanism to learn the feature expression between channels, the proposed method leads to the best improvement in terms of four metrics. It learns from the magnitude spectrum and phase spectrum of the individual microphone signals and exploits the difference in the spatial characteristics of the speech and noise sources.

In the four kinds of microphone array structures designed in the experiment, we obtain 13.60 dB improvement in SI-SNR in the structure of 2 microphones with −10 dB SNR and 14.76 dB improvement in SI-SNR in the distributed 4-microphone array structure.

Another conclusion from the experimental results is that the array structure with four microphones is better than that with the two microphones, indicating that the more the channels are, the more the feature information can be provided to the speech enhancement model. In addition, compared with the other structures, the 4-channel distributed microphone array has the optimum performance. The SDR increase [15.24, 13.87], respectively, in SNR = −10 dB and −5 dB, and the performance improvement of the other

Table 1: Evaluation results of our proposed model compared with other methods on the same dataset. Four metrics and two SNRs are considered.

Method	# Mics	SDR		SI-SNR		PESQ		STOI	
		−10 dB	−5 dB	−10 dB	−5 dB	−10 dB	−5 dB	−10 dB	−5 dB
Noise	2 linear	−6.72	0.34	−7.19	0.24	1.07	1.13	0.40	0.64
	4 linear	−5.93	0.11	−6.40	0.02	1.06	1.11	0.40	0.64
	2 × 2 dB	−6.72	0.32	−6.18	0.22	1.07	1.13	0.40	0.64
	4 dB	−6.89	0.68	−7.17	0.60	1.07	1.13	0.46	0.70
+NCC	2 linear	7.45	11.98	6.42	11.30	1.20	1.76	0.73	0.87
	4 linear	8.89	12.44	7.86	11.75	1.28	**1.83**	**0.75**	**0.88**
	2 × 2 dB	7.63	13.14	7.50	12.37	1.23	1.85	0.71	0.89
	4 dB	8.38	14.58	7.60	13.96	1.30	2.02	**0.76**	**0.91**
+ICD	2 linear	7.42	11.95	6.39	11.27	1.19	1.75	0.72	0.86
	4 linear	8.80	12.38	7.79	11.68	1.27	1.82	0.74	0.87
	2 × 2 dB	7.55	13.06	7.44	12.30	1.22	1.83	0.71	0.89
	4 dB	8.33	14.51	7.56	13.93	1.27	2.01	**0.76**	**0.91**
+MCS	2 linear	7.40	11.92	6.36	11.25	1.19	1.75	0.72	0.86
	4 linear	8.83	12.40	7.83	11.72	1.27	1.82	0.74	0.87
	2 × 2 dB	7.52	13.04	7.42	12.28	1.21	1.82	0.70	0.88
	4 dB	8.28	14.46	7.52	13.79	1.25	2.00	0.75	0.90
Proposed	2 linear	**7.52**	**12.06**	**6.48**	**11.36**	**1.22**	**1.77**	**0.74**	**0.88**
	4 linear	**8.94**	**12.49**	**7.91**	**11.83**	**1.29**	**1.83**	**0.75**	**0.88**
	2 × 2 dB	**7.72**	**13.23**	**7.58**	**12.45**	**1.24**	**1.86**	**0.72**	**0.90**
	4 dB	**8.41**	**14.66**	**7.66**	**14.07**	**1.33**	**2.04**	**0.76**	**0.91**

The best evaluation results are shown in bold, comparing the results of the four speech enhancement methods used in the four microphone arrays.

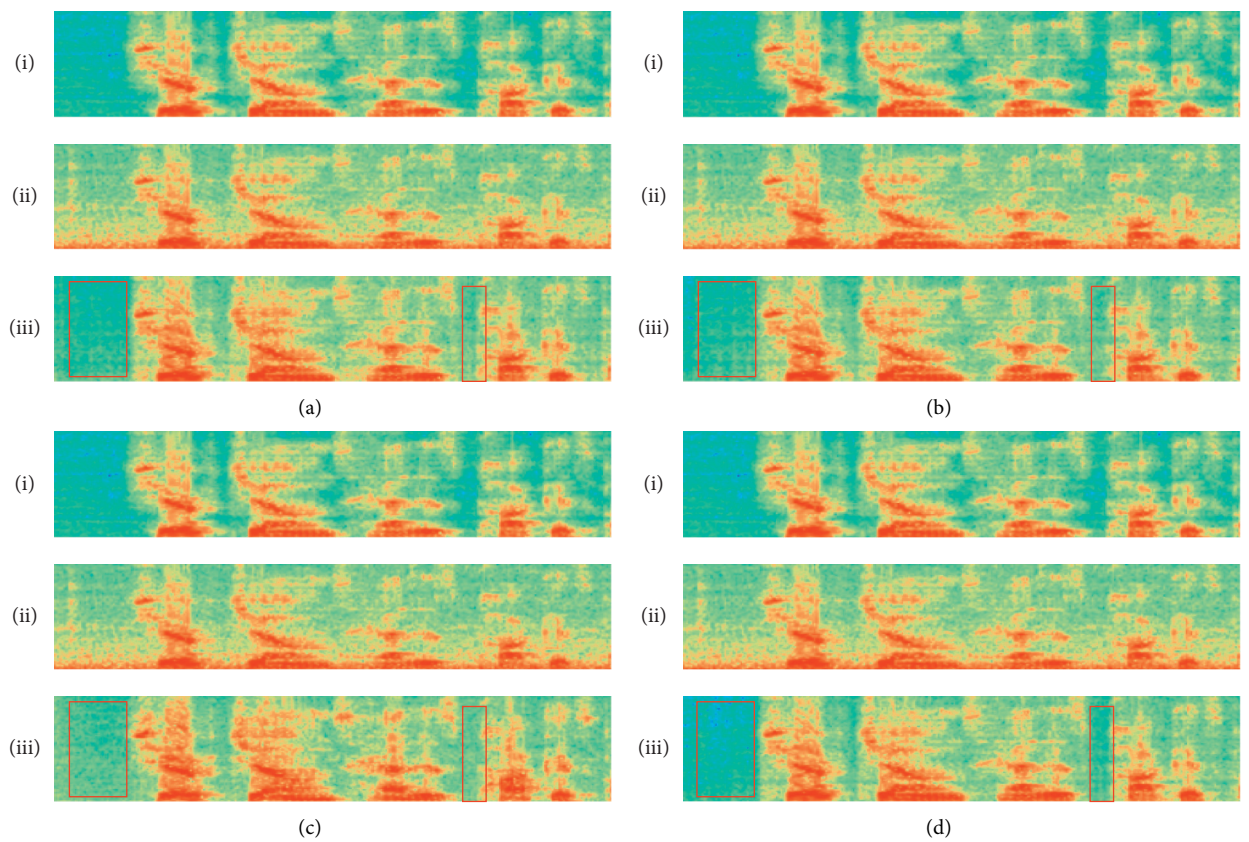


Figure 6: The spectrum of four methods under distributed 4-channel microphone array. (a) NCC method. (b) ICD method. (c) MCS method. (d) IAF method, where (i) represents clear speech, (ii) represents noisy speech with SNR = −10 dB, and (iii) represents denoised speech.

structures are [14.16, 11.63], [14.79, 12.31], [14.32, 12.79], respectively. The distributed microphone array structure has advantages in obtaining the spatial characteristics of the entire cockpit due to the difference in the location of the speech source and the noise source, which is helpful to train a better beamforming filter.

Figure 6 is the speech spectrogram, including the pure speech spectrogram, the noisy speech spectrogram with SNR = −10 dB, and the speech spectrogram enhanced by four methods. The four methods have good noise reduction effects. Compared with the enhanced noise energy spectrum in the box, the method proposed in this paper has significant advantages. At the same time, compared with the enhanced speech spectrogram and pure speech spectrogram, the method did not cause speech damage and ensured the integrity of the speech signal.

5. Conclusions

This work proposed an interchannel attention mechanism frame by frame (IAF) method and jointed with the two-stage BiLSTM network to learn the spatial features directly from multichannel waveforms to solve the problem of multichannel speech enhancement in the car cockpit. Experimental results show the IAF method is more effective than the traditional NCC, MCS, and ICD method in learning spatial features directly from the multichannel speech waveforms. The proposed model based on four distributed microphone arrays obtains the optimal enhancement performance in terms of SDR, SI-SNR, STOI, and PESQ. The results indicated that the method is suitable for different structures of the microphone array and has good robustness. This work provided valuable conclusions for improving the performance of multichannel speech enhancement in the vehicle cockpit. In future work, we will explore the effect of the position of the voice source on the performance using the proposed method.

Data Availability

In order to facilitate the further research of other researchers, the LibriSpeech data in this article can be found at http://www.openslr.org/12/, the Volvo car noise data can be found at http://spib.linse.ufsc.br/noise.html, and the spsquare noise data can be found at https://zenodo.org/record/1227121#.YP0sjo4zZhG.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the fund project of Education Department of Liaoning Province (nos. LJKZ0338 and LJ2020FWL001) and the Undergraduate Innovation and Entrepreneurship Training Project (no. 202110147019).

References

[1] P. Lei, M. Chen, and J. Wang, "Speech enhancement for in-vehicle voice control systems using wavelet analysis and blind source separation," *IET Intelligent Transport Systems*, vol. 13, no. 4, pp. 693–702, 2019.

[2] M. Vollrath, "Speech and driving - solution or problem?" *IET Intelligent Transport Systems*, vol. 1, no. 2, pp. 89–94, 2007.

[3] S. Chakrabarty and E. A. P. Habets, "Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 787–799, 2019.

[4] Y. Masuyama, M. Togami, and T. Komatsu, "Consistency-aware multi-channel speech enhancement using deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 821–825, Barcelona, Spain, May 2020.

[5] P. Tzirakis, A. Kumar, and J. Donley, "Multi-channel speech enhancement using graph neural networks," in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Ontario, Canada, June 2021, http://arxiv.org/abs/2102.06934.

[6] Y. Luo, E. Ceolini, C. Han, C. Enea, and S.-C. Liu, "FaSNet: low-latency adaptive beamforming for multi-microphone audio processing," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 260–267, Singapore, December 2019.

[7] V. Kuzmin, F. Kravchenko, A. Sokolov, and G. Jie, "Real-time streaming wave-u-net with temporal convolutions for multichannel speech enhancement," *Signal Processing*, https://arxiv.org/abs/2104.01923, 2021.

[8] N. Tawara, T. Kobayashi, and T. Ogawa, "multi-channel speech enhancement using time-domain convolutional denoising autoencoder," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association: Crossroads of Speech and Language, INTERSPEECH 2019*, pp. 86–90, Graz, Austria, September 2019.

[9] R. Gu, S. X. Zhang, L. Chen et al., "Enhancing end-to-end multi-channel speech separation via spatial feature learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7319–7323, Barcelona, Spain, May 2020.

[10] S. Araki, T. Hayashi, M. Delcroix, F. Masakiyo, T. Kazuya, and N. Tomohiro, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 116–120, South Brisbane, QLD, Australia, April 2015.

[11] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, & Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[12] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 375–378, Munich, Germany, April 1997.

[13] V. Panayotov, G. Chen, D. Povey, and K. Sanjeev, "Librispeech: an ASR corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, South Brisbane, QLD, Australia, April 2015.

[14] A. Varga, H. Steeneken, and M. Tomlinson, "The NOISEX–92 study on the effect of additive noise on automatic speech

recognition," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1992, technical report speech research unit defense research agency.

[15] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: a database of multichannel environmental noise recordings," *Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3597, 2013.

[16] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "GPURIR: a python library for room impulse response simulation with GPU acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 1–19, 2021.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[18] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, pp. 5998–6008, Long Beach, CA, USA, December 2017.