IEICE TRANSACTIONS

on Information and Systems

DOI:10.1587/transinf.2024EDL8014

Publicized:2024/08/26

This advance publication article will be replaced by the finalized version after proofreading.

A PUBLICATION OF THE INFORMATION AND SYSTEMS SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3 chome, Minato-ku, TOKYO, 105-0011 JAPAN

LETTER

CNN-based feature integration network for speech enhancement in microphone arrays

Ji XI[†], **Pengxu JIANG**[††], ***Nonmembers***, **Yue XIE**[†††], ***Member***, **Wei JIANG**[†], ***and*** **Hao DING**[†], ***Nonmembers***

SUMMARY The relevant model based on convolutional neural networks (CNNs) has been proven to be an effective solution in speech enhancement algorithms. However, there needs to be more research on CNNs based on microphone arrays, especially in exploring the correlation between networks associated with different microphones. In this paper, we proposed a CNN-based feature integration network for speech enhancement in microphone arrays. The input of CNN is composed of short-time Fourier transform (STFT) from different microphones. CNN includes the encoding layer, decoding layer, and skip structure. In addition, the designed feature integration layer enables information exchange between different microphones, and the designed feature fusion layer integrates additional information. The experiment proved the superiority of the designed structure.

key words: *Speech enhancement, convolutional neural network, microphone arrays, deep learning.*

1. Introduction

Multi-microphone noise reduction technology refers to reducing the impact of environmental noise on speech signals through the collection and signal processing of multiple microphones, thereby improving the quality of speech communication [1]. Traditional single microphones are often subject to various environmental noise interferences when collecting speech, such as human voices, car sounds, wind sounds, etc. These noises can cause speech signal distortion and reduce speech recognition accuracy [2]. Multi-microphone noise reduction technology can eliminate or reduce the impact of these noises by fusing and processing signals from multiple microphones.

Deep learning based multi-microphone noise reduction is a technique that uses neural network algorithms to process audio signals recorded by multiple microphones to reduce noise interference [3]. It efficiently and accurately removes environmental noise by extracting valuable speech information from complex noise using deep learning models. Currently, deep learning based multi-microphone noise reduction technology has been widely applied [4-6].

The conventional speech denoising model [7-8] in deep learning typically comprises an encoder and a decoder. Encoding involves transforming the noisy signal, whereas the decoding process aims to recover a clean speech signal by utilizing the received information. Furthermore, it is typical for the encoder and decoder to be interconnected via a skip structure. When dealing with the task of reducing noise using numerous microphones, it is common practice to combine data from many microphones as input for a single model. However, this could lead to the system's inability to retrieve independent information for various microphones. One approach involves establishing separate networks for individual microphones. However, this configuration may result in the loss of correlation information among the different microphones.

In this paper, we designed a convolutional neural network (CNN) based feature integration network for speech enhancement in microphone arrays. The main structure of the model is shown in Fig. 1. Short-time Fourier transform (STFT) is the model input, and CNN is used to obtain time-frequency related information of STFT. In addition, the CNN network consists of encoder and decoder layers [9] and includes a skip structure designed for symmetric encoders. In order to enhance the acquisition of feature-related information across various microphones, the designed model is mainly divided into two paths, One of the paths is used to learn different microphone features separately, including $X_1 \in R^{T\times F\times 1}$, $X_2 \in R^{T\times F\times 1}$,..., $X_n \in R^{T\times F\times 1}$, where n is the number of microphones, T corresponds to the time dimension, F corresponds to the frequency dimension. Another path is used to learn the combination X_A of all microphone features, where $X_A \in R^{T\times F\times n}$. Due to the need to learn the STFT of multiple microphones for noise reduction tasks, CNN may lose associated time-frequency information when learning different features separately. Given this, we have designed a feature integration layer to replace the original skip mechanism. The feature integration layer can gather weighted information from all microphones and provide feedback to their channels. In addition, a feature fusion layer was devised in order to integrate features utilizing weight calculation of output information across several microphones. Finally, the outputs of the two paths are fused to fit the corresponding STFT of pure speech.

[†]The authors are with School of Computer Information Engineering, Changzhou Institute of Technology, Changzhou , 213022, P.R.China.

[††]The author is with the School of Information Science and Engineering, Southeast University, Nanjing, 210096, P.R. China.

[†††]The authors are with School of Information and Communication Engineering, Nanjing Institute of Technology, Nanjing211167, P.R.China.

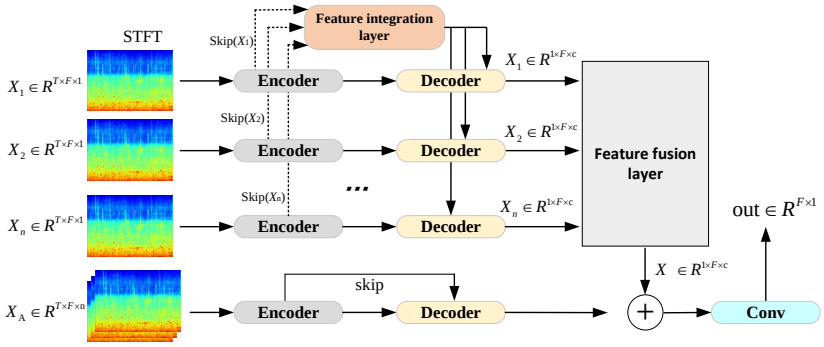

Fig. 1 Illustration of the proposed model.

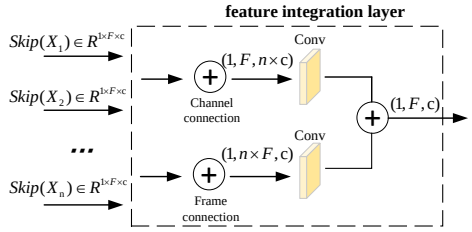

Fig. 2 Illustration of the feature integration layer.

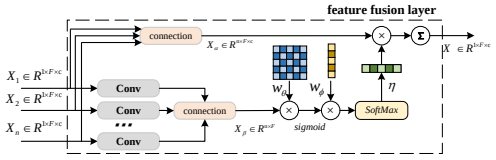

Fig. 3 Illustration of the feature fusion layer.

2. System description

2.1 Convolutional neural network

The designed baseline CNN consists of an encoder and a decoder [10]. The encoder consists of a complex number of lower sampling layers, batch normalization layer, and activation function layer, and the decoder consists of an upper sampling layer, batch normalization layer, and activation function layer. In addition, two jump structures are included for each convolutional network to fit the symmetric convolutional encoder. The encoding and decoding layers in two CNN paths have the same parameter shape.

2.2 Feature integration layer

The feature integration layer is used to connect the encoder layer and decoder layer of CNN, to replace skip structure. The structure of the feature integration layer is shown in Fig. 2. $Skip(X_1)$, $Skip(X_2)$,..., $Skip(X_n)$ are used as inputs. Firstly, perform channel connection and frame connection on all input features. Then, frame and channel convolution are used to obtain corresponding information for different inputs, respectively. Specifically, the convolution kernel size for frame convolution is [1, k], and for channel convolution is [1,1], where k corresponds to the frame dimension. Finally, the feature fusion of different paths serves as the output of the module. In addition, each CNN contains two feature integration layers, corresponding to convolutional layers of different depths.

2.3 Feature fusion layer

The feature fusion layer integrates all features using the weight distribution between input tensors. The structure of the feature fusion layer is shown in Fig. 3. The input of the feature fusion layer is the training results of the microphone array in CNN. All input tensors first pass through a convolutional layer with a shape and number of kernels of 1 to reduce the original input parameters. Then, concatenate all input features to form a feature set $X_\beta \in R^{n\times F}$, where n is the number of microphone arrays, and F corresponds to the frequency dimension. Subsequently, two dense layers map the input features to the specified space and obtain the weight coefficients between different microphones using the $SoftMax$ function:

$$y = \sigma(X_\beta W_\theta)W_\phi, \tag{1}$$

$$\eta = SoftMax(y) \in R^{n\times 1}, \tag{2}$$

here $W_\theta \in R^{F\times F}$, $W_\phi \in R^{F\times 1}$, σ is the activation function *sigmoid*. Finally, concatenate the time dimensions of all input features, which can be represented as X_α, multiply and accumulate them with the corresponding η, as the output X of the feature fusion layer.

3. Experiments

3.1 Preprocessing

Dataset : We utilize CHiME3[11] dataset to show the per-

formance of our proposed model. CHiME3 was developed as part of The 3rd CHiME Speech Separation and Recognition Challenge. We selected isolated 7138 English speech samples for the pure speech of our model, Using four types of noise (Cafe, Street, bus, Pedestrian) as our noise samples to generate noisy speech randomly. All data is provided as 16-bit WAV files sampled at 16kHz. The training set accounts for approximately 80%.

In the simulation experiment, the far-field model linear microphone array is used and placed in a room acoustic environment of 6×5×3m. The coordinates of the center of the microphone array are (2, 3, 1.5), and the distance between adjacent microphones is 0.02m. The coordinates of the three microphones are (2.02, 3, 1.5), (2, 3, 1.5) and (1.98, 3, 1.5). The room reverberation environment is realized through the IMAGE [12] algorithm based on the Allen and Berkley image algorithm, and the reverberation time is 300ms. The sampling rate of the speech signal is set to 16 kHz. The target sound source is located 1m away from the center of the microphone array, and the incident direction is 90°. The interference source is 16kHz white noise from the NOISEX-92 noise database. The interference source is about 1.5m away from the array's center, located in a 180° direction, and the signal-to-interference ratio is set to 40dB. In such an acoustic simulation environment, a multi-mic speech dataset is generated by inputting different single-channel target speech signals. Therefore, we obtained three additional microphone inputs, totaling four.

Feature Generation :To obtain STFT, we defined a periodic Hamming window with a length of 256 and a hop count of 64, removing the symmetric half to obtain the top 129 points. In addition, our input consists of the current STFT noise vector plus the previous seven noise STFT vectors, which means that the input size of one vector is (129,8,1).

Model parameter :The baseline CNN we use is mainly based on [9]. The model parameters are trained through the Adam optimizer, with a batch size of 512 for each training session and a learning rate of 0.0001. The detailed parameters of the baseline CNN are shown in Table 1. "Conv" is the convolutional layer.

Evaluation Metric : Short-time Objective Intelligibility (STOI)[13] and Perceptual Evaluation of Speech Distortion (PESQ) [14] are used to evaluate the designed model.

3.2 Experiment

The main contribution of this article is to propose a CNN structure for multi-microphone noise reduction, and based on this structure, propose a feature integration layer and a feature fusion layer. To verify the structure and related algorithms proposed in this article, Table 2 shows the denoising experiment for multiple microphones, The following is an introduction to different experimental strategies.

- CNN-A: X_A as the input of the baseline CNN. only includes the bottom path in Fig. 1.

Table 1 Proposed baseline CNN.

	Modules	Description
Encoder	Padding	(4,4), (0,0)
	Conv_1	Kernal:$9 \times 8, 18$
	Conv_2	Kernal:$5 \times 1, 30$
	Conv_3	Kernal:$9 \times 1, 8$
	Conv_4	Kernal:$9 \times 1, 18$
	Conv_5	Kernal:$5 \times 1, 30$
	Conv_6	Kernal:$9 \times 1, 8$
Decoder	Conv_7	Kernal:$9 \times 1, 18$
	Conv_8	Kernal:$5 \times 1, 30$
	Conv_9	Kernal:$9 \times 1, 8$
	Conv_10	Kernal:$9 \times 1, 18$
	Conv_11	Kernal:$5 \times 1, 30$
	Conv_12	Kernal:$9 \times 1, 8$
	Dropout	0.2
	Conv_13	Kernal:$129 \times 1, 1$

Table 2 Results of multiple microphones system.

module	PESQ	STOI
CNN-A	1.420	0.780
CNN-B	1.680	0.830
CNN (w/ FF)	1.640	0.857
CNN (w/ FI)	1.534	0.850
CNN MM	1.836	0.858

- CNN-B: The CNN structure proposed in this article. As shown in Fig. 1, excluding the feature integration layer and feature fusion layer, the skip structure connects the encoder layer and decoder layer.
- CNN (w/FF): Including the proposed CNN architecture and feature fusion layer.
- CNN (w/FI): Including the proposed CNN architecture and feature integration layer.
- CNN MM: As shown in Fig. 1, including all proposed modules.

We can conclude from the observation data that the PESQ and STOI are improved by the proposed CNN structure. Firstly, compared to CNN-A, CNN-B's PESQ and STOI have increased by 0.26% and 0.05%, indicating the necessity to consider both individual microphone information and comprehensive microphone information simultaneously. In addition, CNN (w/FF) and CNN (w/FI) can further improve the values of PSEQ and STOI, indicating the effectiveness of the proposed module. CNN MM achieved the best performance, indicating the superiority of the noise reduction architecture proposed in the paper for multiple microphones.

In order to further explore the performance of our designed model, we analyzed the denoising effects of CNN-A and CNN MM in both the time and frequency domains. Fig. 4 shows the denoising results of different models after adding noise to the original speech. Fig. 5 compares the noise reduction effects of different models in the frequency domain under different noise environments. Among them, 'BUS',' CAF ',' PED ', and' STR 'refer to different noise environments' On the bus',' Cafe ',' Pedestrian area ', and' Street ', respectively. From the waveform of the denoised

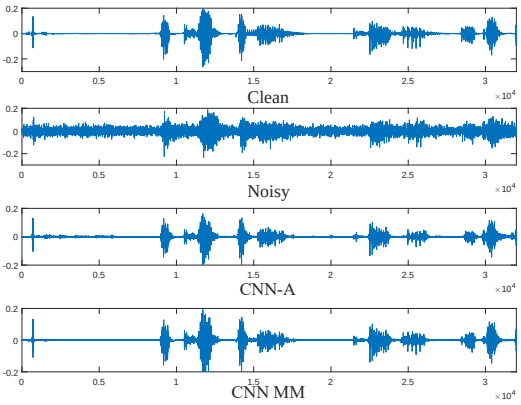

Fig. 4 Waveform samples.

Table 3 Comparison of different methods.

module	PESQ	STOI
DDAEC[7]	1.546	0.852
CRN[8]	1.626	0.844
ours	1.836	0.858

audio, we can see that for silent segments, CNN MM can more effectively eliminate environmental noise than CNN-A, which proves the effectiveness of the proposed feature integration layer and feature fusion layer. In addition, it can be more clearly seen from the noise reduction spectra of different noises that CNN greatly improves its denoising performance at low frequencies, especially in "CAF" and "STR". CNN MM can eliminate more irrelevant information in low frequencies, which proves the effectiveness of the proposed architecture.

Next, we compare the proposed method with models with similar structures, including DDAEC [7] and CRN [8], X_A as input to the model. The comparison of the results of all experiments is shown in Table 3. From the table, even compared to speech enhancement models with similar structures, our proposed model still has performance advantages, which proves the superiority of our proposed multi-microphone speech enhancement model.

4. Conclusion

This paper presented a CNN-based feature integration network for speech enhancement in microphone arrays. STFT is the model input. CNN with encoder, decoder, and skip structure as a baseline model. In addition, we designed a feature integration layer in a multi-microphone-based CNN path to replace the original skip structure and a feature fusion layer to fuse different microphone information. Multiple experimental results have demonstrated the superiority of our designed model.

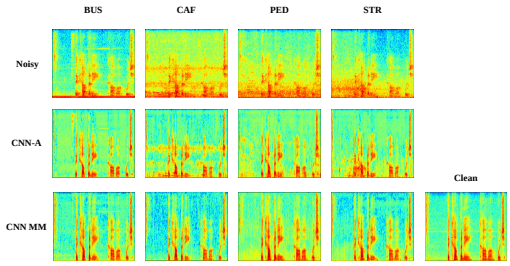

Fig. 5 Spectrogram samples.

References

[1] Das, Nabanita, et al. "Fundamentals, present and future perspectives of speech enhancement." International Journal of Speech Technology 24 (2021): 883-901.
[2] Vihari, Siddala, et al. "Comparison of speech enhancement algorithms." Procedia computer science 89 (2016): 666-676.
[3] Cui, Xingyue, Zhe Chen, and Fuliang Yin. "Multi-objective based multi-channel speech enhancement with BiLSTM network." Applied Acoustics 177 (2021): 107927.
[4] Taherian, Hassan, et al. "Robust speaker recognition based on single-channel and multi-channel speech enhancement." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 1293-1302.
[5] Tan, Ke, Xueliang Zhang, and DeLiang Wang. "Deep learning based real-time speech enhancement for dual-microphone mobile phones." IEEE/ACM transactions on audio, speech, and language processing 29 (2021): 1853-1863.
[6] Barhoush, Mahdi, et al. "Localization-Driven Speech Enhancement in Noisy Multi-Speaker Hospital Environments Using Deep Learning and Meta Learning." IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2022): 670-683.
[7] A. Pandey and D. Wang, "Densely Connected Neural Network with Dilated Convolutions for Real-Time Speech Enhancement in The Time Domain," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6629-6633.
[8] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement", Proc. Interspeech, pp. 3229-3233, Sep. 2018.
[9] Park, Se Rim, and Jinwon Lee. "A fully convolutional neural network for speech enhancement." arXiv preprint arXiv:1609.07132 (2016).
[10] Pandey, Ashutosh, and DeLiang Wang. "Dense CNN with self-attention for time-domain speech enhancement." IEEE/ACM transactions on audio, speech, and language processing 29 (2021): 1270-1279.
[11] Barker, Jon, et al. "The third 'CHiME'speech separation and recognition challenge: Dataset, task and baselines." 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015.
[12] Lehmann, Eric A., and Anders M. Johansson. "Diffuse reverberation model for efficient image-source simulation of room impulse responses." IEEE Transactions on Audio, Speech, and Language Processing 18.6 (2009): 1429-1439.
[13] Taal, Cees H., et al. "A short-time objective intelligibility measure for time-frequency weighted noisy speech." 2010 IEEE international conference on acoustics, speech and signal processing. IEEE, 2010.
[14] Torcoli, Matteo, Thorsten Kastner, and Jürgen Herre. "Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 1530-1541.