IEICE
TRANSACTIONS

on Information and Systems

DOI:10.1587/transinf.2024EDL8042

Publicized:2024/09/10

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

BP-CRN: A Lightweight Two-Stage Convolutional Recurrent Network For Multi-channel Speech Enhancement*

Cong PANG[†a)], **Ye NI**[†], **Jia Ming CHENG**[†], **Lin ZHOU**[†], ***and*** **Li ZHAO**[†], ***Nonmembers***

SUMMARY In our work, we propose a lightweight two-stage convolutional recurrent network (BP-CRN) for multichannel speech enhancement (mcse), which consists of beamforming and post-filtering. Drawing inspiration from traditional methods, we design two core modules for spatial filtering and post-filtering with compensation, named BM and PF, respectively. Both core modules employ a convolutional encoding-decoding structure and utilize complex frequency-time long short-term memory (CFT-LSTM) blocks in the middle. Furthermore, the inter-module mask module is introduced to estimate and convey implicit spatial information and assist the post-filtering module in refining spatial filtering and suppressing residual noise. Experimental results demonstrate that, our proposed method contains only 1.27M parameters and outperforms three other mcse methods in terms of PESQ and STOI metrics.

key words: *multichannel speech enhancement, lightweight, neural beamforming, convolutional recurrent network, complex network.*

1. Introduction

Multichannel speech enhancement involves utilizing multichannel noisy speech to reconstruct and restore clear speech. As the presence of noise inevitably diminishes speech quality and clarity, speech enhancement has attracted significant interest from researchers in recent years. Compared to single-channel speech, microphone arrays can capture much richer spatial and inter-channel information of the target speech signal, leading to improved performance in tasks such as noise suppression, dereverberation, and speech recognition [1], [2].

As deep neural networks (DNNs) become increasingly prevalent in various audio processing fields, they have proven to be highly effective methods [3], [4]. For multichannel audio processing, there are two primary applications of DNNs. One approach involves combining DNNs with traditional signal processing methods, which is called mask-based beamforming. For mask-based beamforming, the role of DNNs is to provide a more accurate estimation of the spatial statistical information of speech and noise for various data-dependent beamformers such as GEV or MVDR [5].

However, beamforming is a linear spatial filter for each frequency bin [6], and the performance of the mask-based beamforming method is limited by the nature of beamforming. Recently, more attention has been focused on performing spatial filtering implicitly using a full neural network, which is called "All-neural Beamformer". The TaylorBeamformer was first introduced for multichannel speech enhancement in [7], where the restoration process is decomposed into a spatial filter and a residual noise canceller. FaSNet [8] employs a two-stage module to directly estimate time-domain beamforming filters, using Transformed Average Connectivity (TAC) to enable the network to utilize information from all microphones.

In this paper, we introduce a lightweight two-stage convolutional recurrent network (BP-CRN) for multichannel speech enhancement. The first core module, BM, derives local representations from complex spectrum and directional feature, capturing correlations along the frequency and time axes through multiple CFT-LSTM modules and implementing spatial filtering. Concurrently, inter-module masking is estimated to better assist the PF module performing further spatial filtering on low-frequency features and post-filtering.

2. Methodology

Fig. 1 presents the overall block diagram of our proposed model. The primary objective is to extract the target speech from the multichannel mixture signal y_c captured by the microphone array, where c denotes the microphone index. Correspondingly, $Y_c(t, f)$ denotes the complex spectrum of the mixture where as c, t, f are the microphone, frame and frequency index, respectively. In the first stage (BM module), the complex spectrum and directional features of the original multichannel mixture signal are fed as the input. A stacked complex FT-LSTMs (CFT-LSTM) proposed in [9] is employed to capture the correlation along the frequency and time axes. Moreover, in the inter-module masking path, the group attention mechanism extracts implicit spatial information from the encoder feature stream, assisting the PF module in further implementing spatial filtering. In the second stage (PF module), a similar encoder-decoder architecture is utilized in conjunction with the inter-module mask to conduct further spatial filtering and suppress residual noise more effectively.

2.1 Directional feature extraction

To better utilize the spatial information present in the original multichannel mixture signal and achieve an improved spatial

[†]The authors are with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China

[*]This work was supported in part by the Key Research and Development Program of Jiangsu Province under Grant BE2022059-3 and the National Natural Science Foundation of China under Grant U2003207.

a) E-mail: pangcong@seu.edu.cn

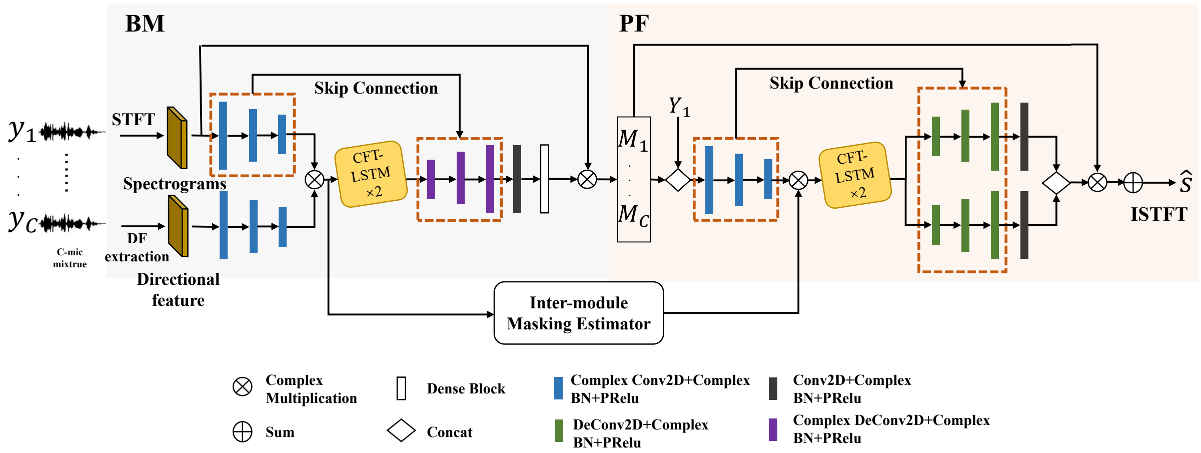

Fig. 1 The network architecture of our proposed BP-CRN model.

filtering effect, we consider extracting directional feature described in [10] to be used as input together with spectral. The inter-channel phase differences (IPD) between p-th pair of microphones can be calculated using the following formula:

$$IPD^{(p)}(t,f) = \angle Y^{p_1}(t,f) - \angle Y^{p_2}(t,f), \quad (1)$$

where, $p = (p1, p2)$ represents the index of the microphone pair, $Y(t,f)$ denotes the complex-valued spectrogram at time t and frequency t, and $\angle(*)$ denotes the phase of spectrogram. Directional feature reveals the similarity between the IPD and the target phase differences (TPD) of each candidate direction, and can better reveal the arrival direction of the sound source in the mixture signal. TPD represents the theoretical phase difference at frequency f between the p-th pair of microphones in the θ direction. Given the microphone array structure, microphone pair index p and azimuth angle θ, we can calculate the TPD in each direction by the following formula:

$$\begin{array}{l} TPD^{(p)}(\theta,f) = \frac{2\pi f}{c} f_s d_p \\ d_p = \Delta_p \cos\theta \end{array}, \quad (2)$$

where, c represents the speed of sound, Δ_p represents the spatial distance between the p-th pair of microphones, fs represents the sampling rate. Directional feature at all candidate azimuths can be calculated as:

$$DF(\theta_i,t,f) = \sum_p \left\langle \mathrm{K}^{IPD^{(p)}(t,f)}, \mathrm{K}^{TPD^{(p)}(\theta,f)} \right\rangle, i = 1,2,...,M, \quad (3)$$

where, $\langle\cdot\rangle$ represent the inner product, $\mathrm{K}^{(*)} = \begin{bmatrix} \cos(*) \\ \sin(*) \end{bmatrix}$ is a 2-D vector composed of the cosine and sine values of the phase difference, and M represents the number of candidate azimuth angles. The larger the value of $V(\theta_i,t,f)$, the greater the probability that there is target speech from the direction θ_i in the mixture signal.

2.2 Beamforming module

The beamforming module (BM) primarily consists of a complex convolutional encoder, a complex convolutional decoder, and the stacked CFT-LSTM blocks in the middle. The obtained directional features are used as the input of the first-stage encoder, and its encoded output is used to weight the encoded multichannel spectrum, corresponding to the complex multiplication operation in the BM block diagram in Figure 1. The encoder comprises three complex convolutional layers, while the decoder features three complex transposed convolutional layers. The 2-D convolutional layer is employed to extract local patterns from the noisy spectrum and reduce feature resolution. In contrast, the decoder utilizes a transposed convolutional layer to restore low-resolution features to their original size. The architecture design of CFT-LSTM used is illustrated in Fig. 2.

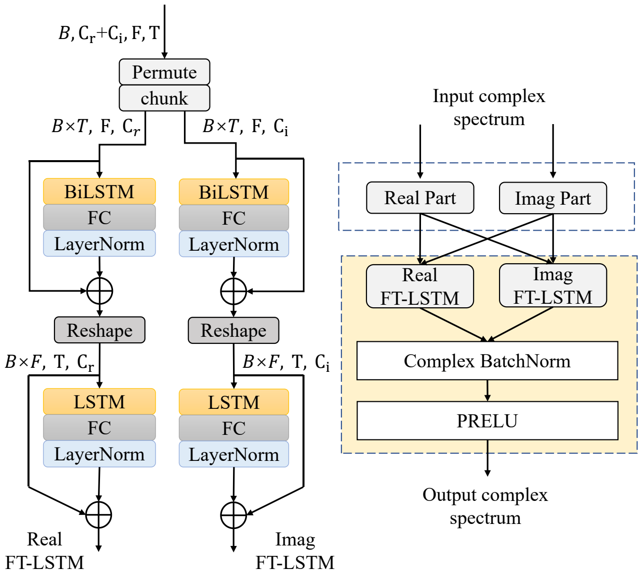

Fig. 2 Architecture design of CFT-LSTM.

2.3 Inter-module mask based on implicit spatial features

We introduce an inter-module mask approximation path be-

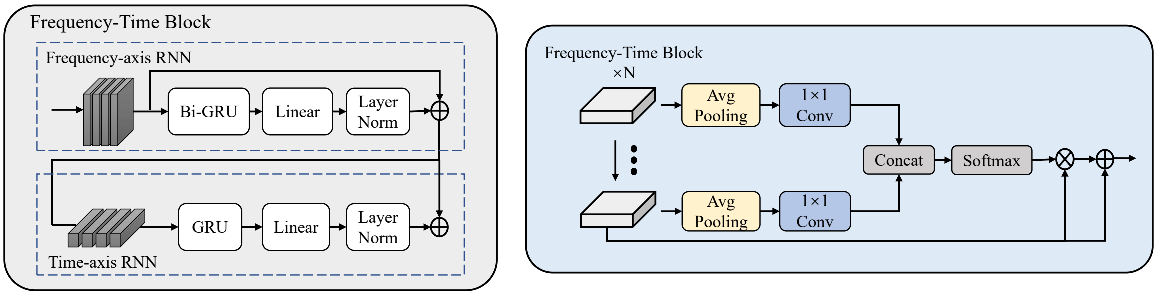

Fig. 3 Architecture design of the inter-module mask approximation path.

tween the spatial filtering module and the post-filtering module. The inter-module mask path employs four frequency-time blocks based on group attention which is illustrated in the Fig. 3. The frequency-axis RNN is employed to model the long-term dependency along the frequency axis, while the time-domain processing block uses the GRU network [11] to perceive long-term correlation in the time domain. Group attention is applied to weight the output of N frequency-time blocks. The inter-module mask extracts features from the encoded output of the spatial filtering module through frequency-time blocks, passing the implicit spatial information to the post-filtering module. This assists the post-filtering module in further spatial filtering and suppressing residual noise.

2.4 The post-filtering module

The main function of the PF module is to perform further spatial filtering and suppress residual noise. we concatenate the spatial filtered signal and the first channel of origin mixture signals and then fed it to the PF module. The purpose of this procedure is to compensate the under-estimated spectral details. After the multichannel input is encoded, it will be element-wise multiplied by the estimated inter-module mask. The inter-module mask extracts the implicit spatial information in the feature stream of the spatial filtering module through the stacked frequency-time blocks and will better assist the post-filtering module performing further spatial filtering on low-frequency features. Then, two separate decoder predict the real and imaginary part of C channel complex ideal ratio mask (cIRM), which will be element-wise multiplied to the STFT of mixture signals to reconstruct the estimation of clean speech spectrum $\hat{S}$.

2.5 Loss function

The model proposed in this paper extracts the STFT coefficients from multichannel noisy speech as input features and the corresponding clean speech as labels, respectively. We train the model by jointly optimizing the mean square error (MSE) [12] of the estimated cIRM and the weighted source distortion ratio loss (Weighted-SDR Loss) [13], with the estimated cIRM as the training target. The model is optimized using a learning rate set to 0.001 via the Adam optimizer. The joint loss function is defined as:

$$\begin{aligned} L_{joint} &= L_{cIRM} + L_{SDR} \\ &= \frac{1}{C}\sum_{n=0}^{C-1} \| \hat{M}{}_r^n - M_r^n \|^2 + \frac{1}{C}\sum_{n=0}^{C-1} \| \hat{M}{}_i^n - M_i^n \|^2 \\ &+ \frac{1}{C}\sum_{n=1}^{C} loss_{wSDR}(x_n, y, \hat{y}_n), \end{aligned} \tag{4}$$

where, C denotes the number of channels, $\hat{M}{}_r^n$ and $\hat{M}{}_i^n$ are the estimated real and imaginary parts of cIRM in channel n, respectively and $loss_{w}SDR$ is the weighted-SDR loss.

3. Experiment

3.1 Experimental setup

We use the publicly available CHIME-3 dataset [14] for training and evaluating speech enhancement performance. The CHIME-3 dataset is a 6-channel (C = 6) microphone recording of talkers speaking in a noisy environment, sampled at 16 kHz. It includes 7138, 1640, 1320 simulation statements for training, development, and test, respectively. We plan to conduct detailed ablation experiments and compare the performance of our proposed algorithm with the current best methods on CHIME-3 dataset. Two evaluation metrics are used: PESQ [15] and STOI [16].

3.2 Ablation experiment

We conducted an ablation experiment on the test set of the CHIME-3 dataset. In ablation experiments, the acronyms 'BM', 'DF', and 'PF' will be utilized to denote the beamforming module, the directional feature extraction module and the postfiltering module, respectively. For fair comparison, all compared models share the same input and labels. By comparing the metrics in the Table 1, it can be seen that for the four noises, every structure proposed is crucial.

3.3 Performance comparison of different algorithms on CHIME-3

To evaluate the effectiveness of different multichannel speech enhancement methods, we select the following four methods to compare with our proposed method (Proposed): the

Table 1 Results of Ablation Studies.

	BUS		STR		CAF		PED	
Method	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
Mic-5 Noisy	1.29	0.875	1.24	0.868	1.23	0.849	1.32	0.880
BM	2.50	0.970	2.23	0.960	2.53	0.968	2.55	0.969
BM+DF	2.54	0.968	2.27	0.960	2.57	0.968	2.55	0.965
BM+DF+PF	2.58	0.971	2.34	0.962	2.58	0.970	2.57	0.966
BP-CRN(Prop.)	**2.63**	**0.974**	**2.38**	**0.965**	**2.66**	**0.972**	**2.68**	**0.971**

baseline **(Mic-5 Noisy)**, which selects the data collected by microphone 5 with the highest signal-to-noise ratio; The U-Net based on CA Dense U-Net **(CADUNet)** proposed by Bahareh et al. [17] for MCSE; The Dense frequency-time attentive network **(DeFTAN)** proposed by Dongheon et al. [18] for MCSE; he dual-path dilated convolutional recurrent network with group attention **(DPDCRN)** proposed by Jiaming et al. [19] for L3DAS23 Challenge.

We compare the performance of our method to the above three state-of-the-art methods on CHiME-3 dataset. In order to make a fair comparison, all the evaluations utilize the same experimental setup. The following Table 2 shows the performance comparison of the proposed method with state-of-the-art results on CHIME-3 dataset. Our proposed method outperforms state-of-the-art results on the CHiME-3 speech enhancement task. Another advantage of this model is its smaller parameter size (1.27M) compared to other models. Furthermore, the complexity of our model (13.334G MAC/s) is significantly less than that of following models.

Table 2 Performance comparison of proposed BP-CRN with state-of-the-art results on CHIME-3.

	SIM-DEV		SIM_TEST		Parameter Size	MAC/s
Method	PESQ	STOI	PESQ	STOI		
Mic-5 Noisy	1.27	0.863	1.27	0.870	/	/
CADUNET	2.39	0.963	2.43	0.959	13.33M	35.251G
DeFTAN	2.43	0.962	2.42	0.954	2.53M	42.735G
DPDCRN	2.46	0.973	2.49	0.959	1.64M	21.878G
BP-CRN(Prop.)	**2.57**	**0.973**	**2.61**	**0.968**	**1.27M**	**13.334G**

4. Conclusions

In our work, we introduce a lightweight two-stage convolutional recurrent network (BP-CRN) for multichannel speech enhancement. Experimental results demonstrate that our proposed method contains only 1.27M parameters and outperforms other methods in terms of PESQ and STOI metrics. For future work, we will attempt to improve the effectiveness of the proposed method at lower signal-to-noise ratios.

References

[1] J. Li, Y. Zhu, D. Luo, Y. Liu, G. Cui, and Z. Li, "The pcg-aiid system for l3das22 challenge: Mimo and miso convolutional recurrent network for multi channel speech enhancement and speech recognition," ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.9211–9215, IEEE, 2022.

[2] X. Xu, R. Gu, and Y. Zou, "Improving dual-microphone speech enhancement by learning cross-channel features with multi-head attention," ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6492–6496, IEEE, 2022.

[3] Y. Xu, J. Du, L.R. Dai, and C.H. Lee, "An experimental study on speech enhancement based on deep neural networks," IEEE Signal processing letters, vol.21, no.1, pp.65–68, 2013.

[4] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," arXiv preprint arXiv:2008.00264, 2020.

[5] H. Erdogan, J.R. Hershey, S. Watanabe, M.I. Mandel, and J. Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks.," Interspeech, pp.1981–1985, 2016.

[6] X. Ji, L. Lu, F. Fang, J. Ma, L. Zhu, J. Li, D. Zhao, M. Liu, and F. Jiang, "An end-to-end far-field keyword spotting system with neural beamforming," 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp.892–899, IEEE, 2021.

[7] A. Li, G. Yu, C. Zheng, and X. Li, "Taylorbeamformer: Learning all-neural beamformer for multi-channel speech enhancement from taylor's approximation theory.," Interspeech, pp.5413–5417, 2022.

[8] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6394–6398, IEEE, 2020.

[9] J. Li, A. Mohamed, G. Zweig, and Y. Gong, "Lstm time and frequency recurrence for automatic speech recognition," 2015 IEEE workshop on automatic speech recognition and understanding (ASRU), pp.187–191, IEEE, 2015.

[10] R. Gu, S.X. Zhang, M. Yu, and D. Yu, "3d spatial features for multi-channel target speech separation," 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp.996–1002, IEEE, 2021.

[11] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[12] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," Machine learning, vol.36, pp.105–139, 1999.

[13] H.S. Choi, J.H. Kim, J. Huh, A. Kim, J.W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," arXiv preprint arXiv:1903.03107, 2019.

[14] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime'speech separation and recognition challenge: Dataset, task and baselines," 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp.504–511, IEEE, 2015.

[15] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221), pp.749–752, IEEE, 2001.

[16] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Transactions on Audio, Speech, and Language Processing, vol.19, no.7, pp.2125–2136, 2011.

[17] B. Tolooshams, R. Giri, A.H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense u-net for multichannel speech enhancement," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.836–840, 2020.

[18] D. Lee and J.W. Choi, "Deft-an: Dense frequency-time attentive network for multichannel speech enhancement," IEEE Signal Processing Letters, vol.30, pp.155–159, 2023.

[19] J. Cheng, C. Pang, R. Liang, J. Fan, and L. Zhao, "Dual-path dilated convolutional recurrent network with group attention for multichannel speech enhancement," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.1–2, 2023.