IMPROVING DUAL-MICROPHONE SPEECH ENHANCEMENT BY LEARNING CROSS-CHANNEL FEATURES WITH MULTI-HEAD ATTENTION

Xinmeng Xu[1]   *Rongzhi Gu*[1]   *Yuexian Zou*[1,2,*]

[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

Hand-crafted spatial features, such as inter-channel intensity difference (IID) and inter-channel phase difference (IPD), play a fundamental role in recent deep learning based dual-microphone speech enhancement (DMSE) systems. However, learning the mutual relationship between artificially designed spatial and spectral features is hard in the end-to-end DMSE. In this work, a novel architecture for DMSE using a multi-head cross-attention based convolutional recurrent network (MHCA-CRN) is presented. The proposed MHCA-CRN model includes a channel-wise encoding structure for preserving intra-channel features and a multi-head cross-attention mechanism for fully exploiting cross-channel features. In addition, the proposed approach specifically formulates the decoder with an extra SNR estimator to estimate frame-level SNR under a multi-task learning framework, which is expected to avoid speech distortion led by end-to-end DMSE module. Finally, a spectral gain function is adopted to further suppress the unnatural residual noise. Experiment results demonstrated superior performance of the proposed model against several state-of-the-art models.

Index Terms— dual-microphone speech enhancement, multi-head cross-attention, SNR estimator, spatial cues extraction, channel-independent encoding

1. INTRODUCTION

Speech communication function of mobile devices has been well-designed and widely used as a convenient tool for contacting others due to its portable characteristics. The quality and intelligibility of the received speech can be severely degraded by background noise if the far-end talker is in an adverse acoustic environment. To attenuate background noise, a two-channel microphone array is typically deployed, where a primary microphone is placed on the bottom of a mobile phone and a secondary microphone on the top.

Conventional dual-channel speech enhancement approaches are based on signal processing, and can be divided into two categories, the blind source separation (BSS) approaches [1, 2], and the beamforming [3, 4] approaches. Although these conventional approaches are fast and lightweight, their performance and robustness are not reliable in a complex acoustic environment.

Recently, with the success of deep learning based single-channel speech enhancement [5, 6], dual-channel speech enhancement works have been developed on exploring deep learning approaches with conventional speech enhancement methods. One is that the deep neural network (DNN) is used to enhance each microphone signal separately, after which a beamformer is used to linearly integrate the dual-channel signals [7, 8]. Experimental results show that the DNN-beamformer approach gives better results than conventional approaches and shows robustness in terms of various noisy types and SNR ranges. However, the DNN only learns temporal-spectral features of each channel while ignoring the spatial features of the target speech in the DNN-beamformer approach. In order to sufficiently leverage the spatial features of dual-channel data, several approaches have been proposed, which apply the spatial features, such as interaural phase or intensity difference (IPD, IID) [9], as additional inputs for improving objective intelligibility and perceptual quality. Despite performance improvement by learning spectral features together with spatial features, the mutual relationship between spatial and spectral information is difficult to learn by a simple DNN, which may cause the under-utilization of spatial information.

To overcome the aforementioned limitation, this paper is motivated to design a convolutional recurrent network (CRN), which can separately process each channel for preserving intra-channel features while interchanging information between encoded channels for fully exploiting the spatial information of dual-channel data. For this purpose, this paper proposes a channel-wise encoding structure to process each input channel independently for preserving intra-channel features, and a multi-head cross-attention (MHCA) mechanism to boost network performance by effectively aggregating cross-channel spatial information. In addition, to maintain superior speech quality, the proposed model formulates the
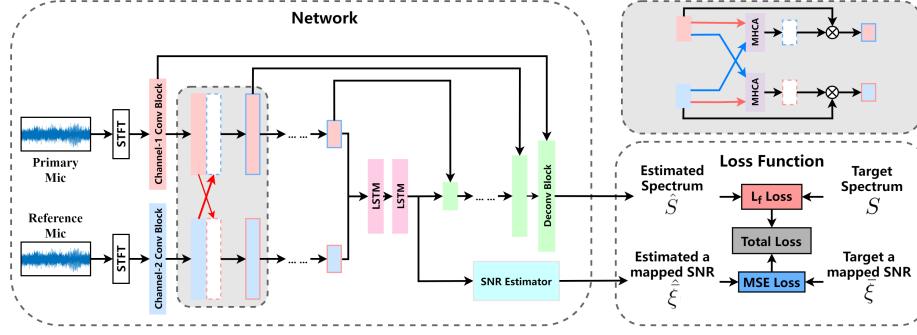
Fig. 1. Schematic diagram of MHCA-CRN. The feature maps generated by the encoding layer are interchanged between channels after processing each convolutional block, which is shown in the gray box. Furthermore, MHCA denotes multi-head cross-attention.

decoder as a multi-task learning framework with an auxiliary task of SNR estimation which has proven to be beneficial to the perceived speech quality [10].

The rest of this paper is organized as follows: the model architecture is presented in Section 2. Section 3 is the dataset and experimental settings. Section 4 demonstrates the results and analysis, and a conclusion is shown in Section 5.

2. PROPOSED MHCA-CRN MODEL

The proposed MHCA based CRN model (MHCA-CRN) treats the dual-microphone enhancement as a supervised learning task, as shown in Fig. 1. First, the proposed model separately encodes the extracted feature from each channel of the noisy signal and interchanges information between encoded channels by using MHCA after each downsampling block. Then the encoded features of both channels are concatenated and fed to LSTM blocks for aggregating temporal contexts. The output of the LSTM blocks is separately fed to the SNR estimator and decoder blocks under a multi-task learning framework. Finally, the output of the SNR estimator block is used to compute a frame-level spectral gain function to remove the residual noise in the estimated spectrum.

2.1. Encoder-decoder structure

The encoder processes each channel independently, in order to preserve the intra-channel feature of dual-channel data and to explicitly utilize the cross-channel relationship. Each encoder contains several stacked 2-D convolutional layers, each of which is followed by batch normalization [11] and exponential linear unit (ELU) [12]. The dilation is applied to the layers along the frequency axis.

The generated feature maps from the encoder of each channel are used as inputs to the MHCA block, which are then interchanged between these two channels. The main objective of the cross-channel attention block is to derive the relationship between two channels.

The decoder is the mirror representation of each encoder except all the convolution layers are replaced with deconvolution layers. Skip connections are introduced to compensate for information loss during the encoding process of the primary microphone.

2.2. Multi-head cross-attention

The MHCA module (shown in Fig 2) is designed for synchronizing time delay between two channels, which holds the spatial information of the target speaker. The MHCA takes transformed feature maps corresponding to the encoder of primary channel, $\mathbf{X}_1$, processed a 1×1-convolution block to form the query, and takes transformed feature maps of the encoder of reference channel, $\mathbf{X}_2$, to form the key-value pair by using two 1×1-convolution blocks. The proposed MHCA first computes the query, and key-value pair for obtaining the attention component $\mathbf{A}$, given by

$$\mathbf{A} = \text{softmax}(\mathbf{Q}\mathbf{K}^\top)\mathbf{V} \tag{1}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the query, key, and value, respectively. Intuitively, the multiplication operation between $\mathbf{Q}$ and $\mathbf{K}$ emphasizes the regions which are slowly varying in time and have high power. What is more, the output of MHCA block $\mathbf{Z}$ is computed by:

$$\mathbf{Z} = \text{sigmoid}(\mathbf{X}_1 + \mathbf{A}) \tag{2}$$

Consequently, the $\mathbf{Z}$ is weight value that are re-scaled between 0 and 1 through a sigmoid activation function.

2.3. SNR estimator

Previous researches have proved that directly training DNN models may inevitably cause a certain amount of speech distortion [13]. To this end, we propose to utilize SNR estimator to estimate frame-level SNR under a multi-task learning framework for maintaining speech quality while reducing noise.
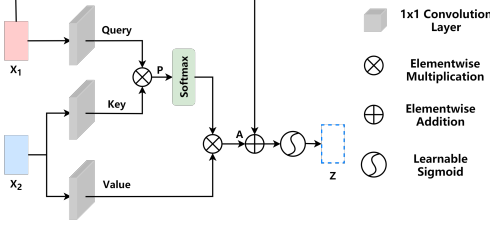
Fig. 2. Multi-head cross-attention module.

Fig. 3. Diagram of target speech reconstruction.

The input of the SNR estimator is the feature maps obtained by two LSTM layers, then it is fed to a convolution layer with sigmoid activation to estimate the frame-level SNR. The training target for the SNR estimator is the mapped SNR [14], which is a mapped version of instantaneous SNR. The definition of instantaneous SNR is as follows:

$$\xi_{dB}(t,f) = 20\log_{10}(|S(t,f)|/|N_1(t,f)|) \qquad (3)$$

where t and f denote the time and frame index, $\xi_{dB}(t,f)$ is scaled in [0, 1] and can be viewed as *a priori* SNR, $S(t,f)$ and $N_1(t,f)$ are respectively the clean and noise spectrum of the primary channel. In addition, it is assumed that $\xi_{dB}(t,f)$ is distributed normally with mean, μ_f, and variance, σ_f^2: $\xi_{dB}(t,f) \sim \mathcal{N}(\mu_f, \sigma_f^2)$. The mapped SNR is given by

$$\bar{\xi}(t,f) = 1 + \left[\frac{1}{2} + \mathrm{erf}\left(\frac{\xi_{dB}(t,f) - \mu_f}{\sigma_f\sqrt{2}}\right)\right] \qquad (4)$$

where "erf" is the error function.

During inference (shown in Fig 3), the SNR estimate, $\hat{\bar{\xi}}(t,f)$ is computed by

$$\hat{\xi}(t,f) = 10^{((\sigma_f\sqrt{2}\mathrm{erf}^{-1}(2\hat{\bar{\xi}}(t,f)-1)+\mu_f)/10)} \qquad (5)$$

where $\hat{\bar{\xi}}(t,f)$ is output from SNR estimator.

2.4. Loss function

Since proposed MHCA-CRN model formulates the decoder with an extra SNR estimator under a multi-task learning framework, the proposed MHCA-CRN model is trained by a combination of two losses. First, the MSE loss to guide the learning of SNR estimator,
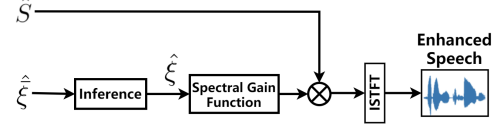
$$L_{SNR} = \mathrm{MSE}(\xi(t,f), \hat{\xi}(t,f)) \qquad (6)$$

Second, the loss [15] for target speech spectrum reconstruction is given by:

$$L_f = \frac{1}{T \times F}\sum_{t=1}^{T}\sum_{f=1}^{F}|(|S(t,f)|) - (|\hat{S}(t,f)|)| \qquad (7)$$

Finally, the total loss is

$$L = L_f + \alpha L_{SNR} \qquad (8)$$

Since both loss values are not on the same scale, we empirically set α to 10.

2.5. Target speech reconstruction

For further suppressing the residual noise, the proposed model adopts the spectral gain function , $G^{SNR}(t,f)$, for dual-microphone speech enhancement, which is represented as

$$G^{SNR}(t,f) = \frac{\hat{\xi}(t,f)}{\hat{\xi}(t,f) + 1} \qquad (9)$$

where $\hat{\xi}(t,f)$ is the estimated SNR obtained by SNR estimator.

The computed final gain $G^{SNR}(t,f)$ is then multiplied to the estimated spectrum $\hat{S}(t,f)$ to suppress the residual noise, which is then combined with the noisy phase to resynthesize the time-domain waveform of the enhanced speech, as shown in Fig 3.

3. EXPERIMENTAL SETUP

3.1. Data preparation

29 hours and 1 hour of speech are selected from Librispeech corpus as training and validation sets, respectively. The noises are from the DEMAND dataset. In addition, this paper simulate room impulse response (RIR) by the IMAGE method [18]. Specifically, two microphones with 2cm interval are placed at the center of a $5m$(length) $\times$ $5m$(width) $\times$ $3m$(height) room, and noise source which are placed at 1.5m away from the center of the two microphones and ranged from $0°$ to $360°$ spaced by $10°$.

For each mixture, a speech and a slice of noise are randomly chosen and are placed at two different positions, and the speech and noise are mixed under the randomly SNR levels ranging from -5dB to 10dB. In addition, the frame length is 32 ms and the hop size is 16 ms. The Hanning window is used as the analysis window. The sampling rate is 16 kHz. A 512-point discrete Fourier transform is used to extract complex short-time Fourier transform (STFT) spectrograms.

3.2. Baselines and training details

The proposed MHCA-CRN model has been compared with four other baselines: (1) **CB-NR** [16]: A coherence-based

Table 1. PESQ and STOI comparison for the different models. A higher score means better performance where the **BOLD** text indicates the best performance for each metric.

Test SNR	Channel	-5 dB		0 dB		5 dB		10 dB	
Meric	-	STOI(%)	PESQ	STOI(%)	PESQ	STOI(%)	PESQ	STOI(%)	PESQ
Unprocessed	Dual	57.06	1.37	69.33	1.88	80.59	2.12	87.63	2.49
DeepXi [14]	Single	77.48	1.88	90.27	2.36	92.17	2.66	95.74	3.11
CB-NR [16]	Dual	54.42	1.46	68.31	2.03	77.57	2.38	88.10	2.74
CRN-PSM [9]	Dual	78.20	1.76	87.30	2.17	92.76	2.59	95.76	2.99
DC-CRN [17]	Dual	86.54	2.48	92.64	2.94	95.88	3.20	97.47	3.43
MHCA-CRN	Dual	**86.58**	**2.51**	**92.83**	**3.03**	**95.96**	**3.24**	**97.52**	**3.46**
-without spectral mapping	Dual	84.98	2.17	89.80	2.57	93.82	2.92	95.47	3.19
-without SNR estimator	Dual	85.76	2.39	92.34	2.88	95.47	3.11	97.02	3.38
-without MHCA blocks	Dual	82.93	2.08	90.06	2.64	94.16	2.98	95.84	3.16
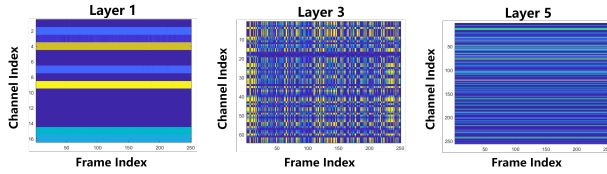
Fig. 4. From left to right: Visualization of MHCA mask at layer 1, layer 3, and layer 5.

dual-channel noise reduction algorithm; (2) **DeepXi** [14]: A minimum mean-square error (MMSE) approach for single-channel speech enhancement by using deep learning; (3) **CRN-PSM** [9]: a CRN approach to predict phase sensitive mask (PSM) for dual-microphone speech enhancement; (4) **DC-CRN** [17]: a densely-connected CRN approach for mobile communication based on dual-channel complex spectral mapping.

To better validate the proposed structure and strategies, we add three ablation experiments. Firstly, we remove the SNR estimator and keep deconvolution layers to predict the clean speech spectrum directly. Secondly, we keep the SNR estimator only and remove the deconvolution layers for comparing the performance between single channel deep learning based MMSE approach, *i.e.* DeepXi, and dual-channel deep learning based MMSE approach. Finally, we remove the MHCA blocks to evaluate the effectiveness of MHCA.

For the training step, all models are trained with Adam optimizer for stochastic gradient descent (SGD) based optimization. The learning rate is set to 0.001. All training samples are zero-padded to have the same number of time steps as the longest sample.

4. RESULTS AND ANALYSIS

The speech enhancement systems are evaluated using Perceptual Evaluation Speech Quality (PESQ) and Short Term Objective Intelligibility (STOI). Experimental results are summarized in Table 1.

According to Table 1, we have the following observa-

tions: Firstly, the deep learning based methods significantly improve both STOI and PESQ metrics, and outperform the conventional approach, *i.e.*, CB-NR. Secondly, the MHCA-CRN without spectral mapping performs better than DeepXi in terms of PESQ and STOI, which indicates that the dual-channel based model show more robustness than single-channel based model. Finally, apart from that, the proposed MHCA-CRN achieves better results than MHCA-CRN without SNR estimator, the proposed MHCA-CRN consistently outperforms the state-of-the-art model, *i.e.*, DC-CRN, in both metrics. This indicate the effectiveness of the SNR estimator.

Table 1 furthermore shows that the MHCA has significantly improved network performance. The visualization of MHCA masks at different convolution layers is as shown in Fig 4. In early layers, the masks pay more attention to certain feature map channels, for example, the 4th and 9th channels of the feature map are highlighted by the mask at the first layer. What is more, when the layer is going deeper, the shape of the MHCA mask is changing in order to synchronize the time delay between two channels. Take, for example, mask at 3rd layer highlights several channels but at different time frames, while mask at 7th highlights certain channels at all time frames. This is demonstrated that the spatial cues between dual-channel can be implicitly exploited by MHCA.

5. CONCLUSION

In this paper, we propose an MHCA-CRN for dual-microphone speech enhancement, aiming to straightforwardly and efficiently exploit spatial information. The model adopts a channel-wise encoding structure to process each input channel independently for preserving intra-channel features and uses the MHCA mechanism to aggregate cross-channel spatial information. Furthermore, an SNR estimator is adopted along with the decoder to estimate frame-level SNR under a multi-task learning framework for further improving the speech quality. Finally, a spectral gain function is adopted to remove unnatural residual noise. Experimental results show that our proposed method can suppress the noise meanwhile maintaining better intelligibility.

6. REFERENCES

[1] Mohamed Djendi and Rédha Bendoumia, "A new adaptive filtering subband algorithm for two-channel acoustic noise reduction and speech enhancement," *Computers & Electrical Engineering*, vol. 39, no. 8, pp. 2531–2550, 2013.

[2] Rahima Henni, Mohamed Djendi, and Mustapha Djebari, "A new efficient two-channel fast transversal adaptive filtering algorithm for blind speech enhancement and acoustic noise reduction," *Computers & Electrical Engineering*, vol. 73, pp. 349–368, 2019.

[3] S Applebaum and Dean Chapman, "Adaptive arrays with main beam constraints," *IEEE Transactions on Antennas and Propagation*, vol. 24, no. 5, pp. 650–662, 1976.

[4] K Buckley and L Griffiths, "An adaptive generalized sidelobe canceller with derivative constraints," *IEEE Transactions on antennas and propagation*, vol. 34, no. 3, pp. 311–319, 1986.

[5] J Rouat, "Computational auditory scene analysis: Principles, algorithms, and applications (wang, d. and brown, gj, eds.; 2006)[book review]," *IEEE Transactions on Neural Networks*, vol. 19, no. 1, pp. 199–199, 2008.

[6] Xinmeng Xu, Yang Wang, Dongxiang Xu, Yiyuan Peng, Cong Zhang, Jie Jia, and Binbin Chen, "Multi-stage progressive speech enhancement network," *Proc. Interspeech 2021*, pp. 2691–2695, 2021.

[7] Xiong Xiao, Shinji Watanabe, Hakan Erdogan, Liang Lu, John Hershey, Michael L Seltzer, Guoguo Chen, Yu Zhang, Michael Mandel, and Dong Yu, "Deep beamforming networks for multi-channel speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5745–5749.

[8] Hakan Erdogan, John R Hershey, Shinji Watanabe, Michael I Mandel, and Jonathan Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks.," in *Interspeech*, 2016, pp. 1981–1985.

[9] Ke Tan, Xueliang Zhang, and DeLiang Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5751–5755.

[10] Aaron Nicolson and Kuldip K Paliwal, "Masked multi-head self-attention for causal speech enhancement," *Speech Communication*, vol. 125, pp. 80–96, 2020.

[11] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[12] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[13] Chengyu Zheng, Xiulian Peng, Yuan Zhang, Sriram Srinivasan, and Yan Lu, "Interactive speech and noise modeling for speech enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 14549–14557.

[14] Aaron Nicolson and Kuldip K Paliwal, "Deep Xi as a front-end for robust automatic speech recognition," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, 2020, pp. 1–6.

[15] Ashutosh Pandey and Deliang Wang, "On adversarial training and loss functions for speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5414–5418.

[16] Youna Ji, Jun Byun, and Young-cheol Park, "Coherence-based dual-channel noise reduction algorithm in a complex noisy environment.," in *INTERSPEECH*, 2017, pp. 2670–2674.

[17] Ke Tan, Xueliang Zhang, and DeLiang Wang, "Real-time speech enhancement for mobile communication based on dual-channel complex spectral mapping," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6134–6138.

[18] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.