ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446389

IMPROVING ACOUSTIC ECHO CANCELLATION BY EXPLORING SPEECH AND ECHO AFFINITY WITH MULTI-HEAD ATTENTION

Yiqun Zhang[1], *Xinmeng Xu*[1], *Weiping Tu*[1,2,*]

[1]NERCMS, School of Computer Science, Hubei Luojia Laboratory,
Wuhan University, China
[2]Hubei Key Laboratory of Multimedia and Network Communication Engineering,
Wuhan University, China

ABSTRACT

Deep learning-based approaches formulate acoustic echo cancellation (AEC) as a supervised speech separation task, where the mixture signal and the far-end signal are combined directly before or after the encoding stage. However, the mixture signal and the far-end signal are not integrated sufficiently due to the lack of interpretability for the affinity between speech and echo in a noisy mixture. In this paper, we propose DCA-Net, a dual-branch cross-attention neural network, to improve AEC performance by exploring the affinities between speech and echo in the representation space. In particular, the two branches predict speech and echo, respectively, and an interaction module is designed at several intermediate feature domains between the two branches to learn the correlations between these features of the two branches. Such an interaction can leverage features learned from one branch to restore missing information or counteract undesired information of the other by calculating the similarity between these features of two branches using multi-head cross attention. Evaluation results show that the proposed DCA-Net effectively suppresses acoustic echo and noise while preserving good speech quality.

Index Terms— Acoustic echo cancellation, multi-head cross-attention, dual-branch, interaction module

1. INTRODUCTION

Acoustic echo is a general problem in full-duplex voice communication scenarios, where the microphone at the near-end picks up audio signals from the loudspeaker and sends them back to the far-end. In this case, the user at the far-end can hear his/her own voice in the loudspeaker signal, which is delayed by the round-trip time of the system and seriously degrades the communication quality. The goal of acoustic echo cancellation (AEC) is to eliminate the echo from the microphone signal while minimizing distortion to the near-end speech [1, 2, 3].

Owing to the capability of modeling complicated nonlinear relationships, deep learning has been utilized recently for addressing AEC problems [4]. Deep learning-based approaches formulate AEC as a supervised speech separation task [5], which separates the echo and the near-end speech from the mixture signal. Zhang et al.[6] directly combine the information of the mixture signal and the far-end signal into a multi-channel feature in the input stage of the model. To effectively suppress the echo through the echo-related information from the far-end signal, Kim utilizes an additional encoder to extract the feature of the far-end signal as auxiliary information for eliminating the echo [7]. What is more, for exploring the potential relationships of the far-end signal and the mixture signal, Sun et al. uses a special encoder in which the features between the far-end signal and the mixture is applied to assist in feature extraction of the mixture signal [8]. Nevertheless, using the echo-related features from the far-end signal as auxiliary information makes it difficult to precisely match the echo feature in the mixture signal and eliminate the echo. The way of directly integrating the feature of the mixture signal and the far-end signal may not fully exploit the correlation between them, causing the lack of interpretability for the affinity between speech and echo in a noisy mixture.

To explore the affinity between speech and echo in the noisy mixture and effectively improve the correlation between the mixture signal and the far-end signal, we propose DCA-Net, a dual-branch cross-attention neural network that predicts near-end speech and echo simultaneously. Particularly, we design an interaction module based on multi-head cross-attention (MHCA) at intermediate feature domains between two branches to capture speech or echo-related features from the mixture signal and the far-end signal. Our interaction can leverage features learned from one branch to restore missing information or counteract the negative impact of undesired information from the other by calculating the similarity be-

[*]Corresponding author

This work was supported in part by the National Nature Science Foundation of China (No. 62071342, No.62171326), the Special Fund of Hubei Luojia Laboratory (No. 220100019), the Hubei Province Technological Innovation Major Project (No. 2021BAA034) and the Fundamental Research Funds for the Central Universities (No.2042023kf1033).

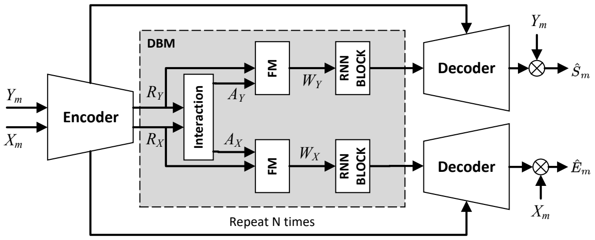

Fig. 1. *Framework of DCA-Net.*

tween these features of the two branches. Moreover, we propose to capture the correlation between processed features by the interaction module and intermediate representations in the fusion module, which ensures that the features extracted from the other branch are positive for the overall system performance.

2. PROPOSED METHOD

2.1. Problem Formulation

The microphone signal $y(n)$ is a mixture of echo $d(n)$, near-end speech $s(n)$, and background noise $v(n)$:

$$y(n) = d(n) + s(n) + v(n) \tag{1}$$

where n is the sample index. The acoustic echo $d(n)$ is formed by the reference signal $x(n)$ through a nonlinear echo path consisting of loudspeaker nonlinear distortions and a room impulse response (RIR) between loudspeaker and microphone $h(n)$. The goal of the AEC is to estimate the clean near-end speech $s(n)$ from the mixture $y(n)$ by jointly suppressing the acoustic echo $d(n)$ and background noise $v(n)$.

2.2. Network Architecture

As illustrated in Fig.1, DCA-Net accepts the magnitude of the short-time Fourier transform (STFT) of $y(n)$ and $x(n)$, Y_m and X_m, and predicts the magnitude of near-end speech and echo, $\hat{S}_m$ and $\hat{D}_m$.

Each branch of DCA-Net is an encoder-decoder architecture, in which the encoder consists of four convolution blocks and the decoder has an equal number of deconvolution blocks. Each convolution block contains a 2D convolution layer, batchnorm(BN) and a PReLU activation function, while deconvolution blocks are the mirror of their corresponding convolution blocks, and skip connections are inserted between the encoder and decoder. The last deconvolution blocks with sigmoid activation produce the magnitude masks for the speech and echo, respectively.

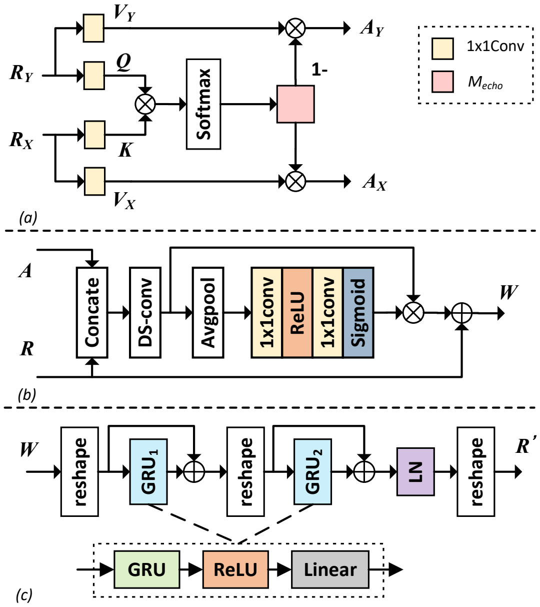

Fig. 2. *Details of DBM: (a) Interaction Module, (b) FM, (c) RNN Block.*

The separator between encoder and decoder is composed of several repeated dual-branch modules (DBM), in which we set up two branches for intermediate representations. Every DBM has a interaction module to extract the speech and echo related features. These features are then combined with their corresponding intermediate representations in a fusion module (FM). Following the fusion module, we utilize a RNN Block for sequence modeling, The comprehensive expositions are presented in 2.3 and 2.4.

About the phase of the target speech, we borrow the idea from [9] to build a phase reconstruction network, which consists of five convolution layers in series. The input is a five channel feature stacked by $\hat{S}_m$ and the real and imaginary parts of the STFT of $y(n)$ and $x(n)$, while the output is the phase of the target speech, $\hat{Y}_\theta$.

The training objective of the cascade architecture contains three parts. We define the loss functions, L_{mag}, the mean squared error (MSE) between $\hat{S}(t, f)$ and $S(t, f)$, as well as L_{echo}, the MSE between $\hat{E}(t, f)$ and $E(t, f)$. The third loss function L_{SI-SNR}, which relates to the speech waveform, is the negative scale-invariant source-to-noise ratio (SI-SNR) [10]. We propose to design a coefficient λ to combine them:

$$Loss = \lambda(L_{mag} + L_{echo}) + L_{SI-SNR} \tag{2}$$

where we empirically select $\lambda = 0.5$ based on the performance on the validation data.

2.3. Interaction Module

Echo features exist in both mixture signals and far-end signals. Inspired by the success of multi-head cross attention(MHCA) mechanism in the dual-microphone speech enhancement task [11], the MHCA mechanism is designed in the interaction module for extracting relevant echo features from both branches, as shown in Fig.2.(a). The MHCA performs linear mapping on the intermediate representations R_X and R_Y to form the query Q, key K, value V_Y and value V_X.

Intuitively, the multiplication operation between Q and K emphasizes the common features of mixture signals and far-end signals. As a result, the part of the echo features outweighs that of the speech-related information. Following the softmax operation, the feature map M_{echo} highlights echo features with high weights.

$$M_{echo} = softmax(QK^T) \tag{3}$$

where Q, $K \in \mathbb{R}^{T\times F\times C}$ and $M_{echo} \in \mathbb{R}^{T\times F\times F}$.

As the weights of speech features are low in comparison to those of echo features, we can obtain M_{speech} highlighting the speech features and suppressing the echo parts by subtracting M_{echo} from 1. Then we utilize matrix multiplication to obtain attention components A_Y and A_X, respectively. The obtained attention components are fed into corresponding branches, where fusion modules integrate them with the intermediate representation.

2.4. Fusion Module and RNN Block

In order to reduce the influence of the feature redundancy caused by residual layers and irrelevant information in encoding representations, we use a fusion module (FM) to selectively aggregate the representations and relevant attention features, and dynamically capture the correlation between them [12].

The depth-wise separable convolutions (DS-Conv) and the channel-wise attention mechanism are incorporated to preserve relevant information and remove redundant features simultaneously in each branch. Both branches share the same structure, as shown in Fig.2.(b). The task of DS-Conv is to aggregate the input representations $R \in \mathbb{R}^{C\times F\times T}$ and attention features A. Then, the channel-wise attention employs a global pooling operation and two 1×1 convolution layers with a ReLU function in between to select information dynamically. The first convolution layer compresses the channel dimension to 1/16 of input, while the subsequent layer reinstates the channel to its original configuration. In addition, a sigmoid activation function is applied to generate a mask for the aggregated features.

After that, the output $W \in \mathbb{R}^{C\times F\times T}$ is fed to RNN Block for sequence modeling. Fig.2.(c) shows this process. W is first reshaped to $\mathbb{R}^{T\times F\times C}$ before being input to GRU_1, which is applied to the frequency dimension. GRU_2 scans the time axis and accepts input with a shape of $\mathbb{R}^{F\times T\times C}$. Finally, the output $R' \in \mathbb{R}^{C\times F\times T}$ is sent to the next DBM or the decoder after the layernorm (LN).

3. EXPERIMENTAL SETUP

3.1. Dataset

The AEC-Challenge dataset [13, 14], an open speech corpus provided by Microsoft, is used during the training stage of our study. In addition, we perform data expansion involving preparing four types of signals: near-end speech, background noise, far-end signal and corresponding echo signal.

We first randomly select 500 utterances of the 2022 official synthetic dataset as the test set which is unseen in training, and choose the remaining 9500 utterances, along with 10000 utterances from 2021 official synthetic dataset, for near-end speech, far-end signal, and corresonding echo signal. For background noise, we randomly select 5000 noise signals from the DNS [15] data. Besides, we also use the real far-end single-talk utterances from the 2021 and 2022 official dataset. Finally, we randomly combine these four types of signals to obtain expanded datasets.

3.2. Evaluation Metrics

Performance of all systems is evaluated in terms of echo return loss enhancement (ERLE) for far-end single-talk periods, perceptual evaluation of speech quality (PESQ) [16] and short-time objective intelligibility (STOI) [17] for double-talk records. The wideband Mean Opinion Score is based on ITU-T Recommendations P.862.2. A higher score means the model performs better.

In this study, ERLE is defined as:

$$\mathrm{ERLE} = 10\log_{10}\,[\sum_n y^2(n)/\sum_n \hat{s}^2(n)] \tag{4}$$

where $y(n)$ is the mixture signal, and $\hat{s}(n)$ is the estimated signal.

3.3. Implementation Details

All training and evaluation audio signals are resampled to 16kHz. Our approach uses STFT with a hann window to extract the spectrum from the utterance. The frame length is 32ms, the hop size is 16ms and the DFT length is 512. In the encoder, the channel number of the convolutional layers is [16, 32, 64, 128]. The kernel size and the stride are respectively set to (5,3) and (2,1) in frequency and time dimension. The two branches share the same encoder. The hidden dimension of GRU is 128 and we use the unidirectional GRU to ensure the causality of the model. The kernel size of phase reconstruction network is [(5, 3), (25, 1), (5, 3), (25, 1), (1, 1)]. DBM repeats 3 times. Our model's parameters are 2.5M.

Table 1. *The objective test results of the baseline model and the proposed model. MOS-LQO of ITU-T P.862.2 ranges between 1.04 to 4.64. DT: doubletalk, ST: single-talk, NE: near-end, FE: far-end, SER: source-to-echo ratio. * indicates that near-end noise is increased at an SNR of [5, 20]dB. Bold results indicate the best in each column.*

Method	DT								ST_NE		ST_FE
SER(in dB)	-5		5		0*		10*		-		-
Metrics	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	ERLE
CRN	1.53	86.7	2.19	94.7	1.57	87.2	1.95	94.0	2.34	95.4	21.2
F-T-LSTM	1.61	88.4	2.32	95.3	1.67	88.6	2.04	94.4	2.51	96.2	25.2
MTFAA	1.69	89.6	2.49	**96.1**	1.74	90.2	2.20	95.2	2.61	96.5	30.1
DCA-Net	**1.81**	**90.1**	**2.62**	96.0	**1.87**	**90.3**	**2.32**	**95.3**	**2.66**	**96.7**	**35.3**

Table 2. *Ablation experiments. The SER in DT scenario is randomly set to [-10, 15]dB. SNR is randomly picked up from [5, 20]dB or without additional noise in DT and is set to [5, 20]dB in ST_NE.*

Method	DT		ST_NE		ST_FE
Metrics	PESQ	STOI	PESQ	STOI	ERLE
DCA-Net	**2.17**	**93.1**	**2.66**	**96.7**	**35.3**
-w/o FM	2.09	92.7	2.60	96.4	32.8
-w/o interaction	2.04	92.2	2.54	95.9	27.9
-w/o RNN Block	2.01	92.0	2.56	96.2	25.5

Prior to training, we employ the GCC-PHAT algorithm [18] to calculate the time delay between the far-end signal and the near-end signal in the time domain. The model is then trained by using the Adam optimizer algorithm on 4-second segments with a learning rate of 2e-4. The batch size is set to 32 for all experiments.

4. RESULTS AND ANALYSIS

We compare our DCA-Net with three other methods, and the experiment results are shown in Table 1, where the bold results are the best. Three baseline models are selected for comparison: (1) CRN: a notable model in speech enhancement, which is widely used in acoustic echo cancellation [19], (2) F-T-LSTM: a novel AEC model that uses frequency-time-LSTMs to scan both frequency and time axis for better temporal modeling the important phase information [6], (3) MT-FAA: a novel AEC model that utilizes a phase encoder, multi-scale time-frequency processing, and streaming axial attention [20]. All models are trained with the dataset described in section 3.1.

According to Table 1, we observe that the proposed DCA-Net outperforms the selected baselines in both double-talk situations and single-talk situations. Obviously, DCA-Net improving the interaction between the mixture signal and the far-end signal can sufficiently utilize the correlation of the two signals to suppress echo. The result in near-end single-talk situations indicates that DCA-Net preserves speech quality and restores the speech while suppressing the noise in the near-end. As a result, we conclude that our proposed model sufficiently exploring the affinity between the speech and echo in the noisy mixture more effectively improves the performance of the AEC.

We also conduct several ablation experiments to evaluate the effectiveness of different model components of DCA-Net in Table 2. (1) DCA-Net-w/o FM: DCA-Net without fusion modules, directly using residual connection to combine attention compenents and encoding representations, (2) DCA-Net-w/o interaction: DCA-Net without interaction and we directly intergrate the feature of the mixture signal and the far-end signal, (3) DCA-Net-w/o RNN Block: DCA-Net without RNN Block.

As shown for the ablation study in Table 2, 0.08 PESQ gains and 0.4 STOI gains by DCA-Net over DCA-Net -w/o FM in double-talk show that the significance of FM to reduce the features redundancy. In addition, in double-talk situation, comparing DCA-Net and DCA-Net-w/o interaction, PESQ and STOI increase by 0.13 and 0.9, respectively. The results demonstrate that our interaction mechanism sufficiently leverages the correlation between the mixture signal and the far-end signal to integrate the feature.

5. CONCLUSION

We propose a dual-branch neural network (DCA-Net) that effectively explores the affinity between speech and echo in a noisy mixture for the AEC task. Particularly, an interaction module base on multi-head cross attention is designed to leverage features learned from another branch to enhance the target signal modeling and counteract negative information. Experimental results show that the interaction module effectually improves the correlation between the mixture signal and the far-end signal compared to the direct combination. The proposed method outperforms other baseline methods by exploring the affinity between speech and echo.

6. REFERENCES

[1] MM Sondhi, "An adaptive echo canceller," *Bell System technical journal*, vol. 46, no. 3, pp. 497–511, 1967.

[2] Jacob Benesty, Tomas Gänsler, Dennis R Morgan, M Mohan Sondhi, Steven L Gay, et al., "Advances in network and acoustic echo cancellation," 2001.

[3] Gerald Enzner, Herbert Buchner, Alexis Favrot, and Fabian Kuech, "Acoustic echo control," in *Academic press library in signal processing*, vol. 4, pp. 807–877. Elsevier, 2014.

[4] Chul Min Lee, Jong Won Shin, and Nam Soo Kim, "Dnn-based residual echo suppression," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] Hao Zhang and DeLiang Wang, "Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios," in *Proc. Interspeech 2018*, 2018, pp. 3239–3243.

[6] Shimin Zhang, Yuxiang Kong, Shubo Lv, Yanxin Hu, and Lei Xie, "F-T-LSTM Based Complex Network for Joint Acoustic Echo Cancellation and Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 4758–4762.

[7] Jung-Hee Kim and Joon-Hyuk Chang, "Attention wave-u-net for acoustic echo cancellation.," in *Interspeech*, 2020, pp. 3969–3973.

[8] Xingwei Sun, Chenbin Cao, Qinglong Li, Linzhang Wang, and Fei Xiang, "Explore relative and context information with transformer for joint acoustic echo cancellation and speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9117–9121.

[9] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 9458–9465, Apr. 2020.

[10] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

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

[12] Kui Jiang, Zhongyuan Wang, Chen Chen, Zheng Wang, Laizhong Cui, and Chia-Wen Lin, "Magic elf: Image deraining meets association learning and transformer," *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 827–836, 2022.

[13] Kusha Sridhar, Ross Cutler, Ando Saabas, Tanel Parnamaa, Markus Loide, Hannes Gamper, Sebastian Braun, Robert Aichner, and Sriram Srinivasan, "Icassp 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 151–155.

[14] Ross Cutler, Ando Saabas, Tanel Parnamaa, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sørensen, and Robert Aichner, "Icassp 2022 acoustic echo cancellation challenge," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9107–9111.

[15] Chandan K.A. Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "INTERSPEECH 2021 Deep Noise Suppression Challenge," in *Proc. Interspeech 2021*, 2021, pp. 2796–2800.

[16] IT Union, "Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, Recommendation P*, vol. 862, 2007.

[17] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[18] Lukas Pfeifenberger, Matthias Zoehrer, and Franz Pernkopf, "Acoustic echo cancellation with cross-domain learning.," in *Interspeech*, 2021, pp. 4753–4757.

[19] Hao Zhang, Ke Tan, and DeLiang Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions.," in *Interspeech*, 2019, pp. 4255–4259.

[20] Guochang Zhang, Libiao Yu, Chunliang Wang, and Jianqiang Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9122–9126.