Interspeech 2024
1-5 September 2024, Kos, Greece

MSA-DPCRN: A Multi-Scale Asymmetric Dual-Path Convolution Recurrent Network with Attentional Feature Fusion for Acoustic Echo Cancellation

Ye Ni[*1], *Cong Pang*[*1], *Chengwei Huang*[2], *Cairong Zou*[1]

[1]Southeast University, China
[2]Zhejiang Lab, China

niye@seu.edu.cn, pangcong@seu.edu.cn, huangcwx@126.com, cairong@seu.edu.cn

Abstract

Echo cancellation plays a crucial role in modern speech applications. Numerous deep-learning models have been developed for the echo cancellation task and achieved great progress by incorporating additional features; however, the majority of these models overlook the characteristics of different features and simply merge them along the channel dimension. In this paper, we proposed a multi-scale asymmetric dual-path convolution recurrent network (MSA-DPCRN) consisting of two asymmetric encoding paths to extract spectrum and relevant features from the input reference and microphone signals. Moreover, we propose a frequency-wise attentional feature fusion (AFF) method to fuse the two features while maintaining the original dynamic range. The experiments validate the effectiveness of each component in MSA-DPCRN and indicate that our model outperforms the AEC challenge baseline in terms of the Echo-MOS metrics.
Index Terms: Speech enhancement, acoustic echo cancellation, attention network, complex network

1. Introduction

Acoustic echo cancellation (AEC) continues to be a significant challenge in various applications, such as hands-free telephones, audio/video conference systems, and hearing aids. The presence of echo in speech poses difficulties in addressing automatic speech recognition, affective computing, and other acoustic post-processing tasks [1, 2].

In deep learning-based AEC methods, echo cancellation is treated as a speech separation task using a convolution recurrent network (CRN) [3] that combines convolution and recurrent layers to capture both local and temporal dependencies in the audio signals. Many of these AEC models utilize a symmetric encoder that can be achieved by either concatenating different features along the channel dimension or encoding input features with two consistent paths [4, 5]. Spectrum features of microphone and reference speech are concatenated as an input in [6, 7, 8, 9, 10]. Besides the spectrum features, outputs of a linear acoustic echo canceller (LAEC) are introduced to the input in [11] to provide adaptive features, and in [12] to help the model identify which time-frequency bin (T-F) regions have been suppressed. Various combinations of microphone, reference, and LAEC outputs have been investigated in [13, 14] to determine the optimal input features for improving performance. These models aim to enhance the input representation by incorporating additional features along the channel dimension.

Despite the advancements made by these models, directly using the same convolution layer for all input features, regardless of their context, can be a bottleneck. Considering that the echo can be regarded as a delayed, scaled, nonlinear version [15, 16] of the reference signal, estimating the echo signal requires incorporating information from previous frames. However, incorporating historical frames in the microphone signal can decrease the spectrum resolution of the current frame. Although there are some research progresses [17, 18] achieved in the field of delay estimation, there are still some challenges in accurately estimating delays. Concerning the distinct characteristics of various features, inconsistent encoding paths are required.

In this paper, we introduce a multi-scale asymmetric dual-path convolution recurrent network (MSA-DPCRN) with attentional feature fusion for echo cancellation. The encoder of the MSA-DPCRN model consists of a spectrum feature extraction path (SFP) and a relevant feature extraction path (RFP), which operates on different temporal scales and handles asymmetric inputs. Inspired by the work [19], we propose a frequency-wise attentional feature fusion (AFF) block to aggregate the features with inconsistent temporal scales. Considering the distinct spectrum distributions between echo and microphone signals, the AFF block is designed to generate fusion weights for selecting features from the outputs of SFP and RFP. The primary contributions of our work can be encapsulated as follows:

(i) A multi-scale asymmetric encoding architecture. The encoding block comprises two distinct paths: SFP and RFP. The SFP is dedicated to extracting spectrogram features from the microphone signal, while the RFP concentrates on extracting relevant features from both the microphone and reference signals. To prioritize the current frame information and minimize interference from previous frames, we establish skip connections between the SFP and decoder, transmitting only the microphone spectrum information.

(ii) An attentional feature fusion block. The AFF is based on the element-wise feature fusion (EAF) approach, which effectively combines and aggregates multi-scale spectrum features. In contrast to EAF methods that treat features from different sources equally, the AFF method introduces a soft selection mechanism that assigns adaptive frequency-wise weights to the features. Additionally, with the adaptive attentional weights, the AFF method overcomes the limitation that EAF would widen the range of feature values and disperse feature distribution. Therefore, the AFF block can provide additional flexibility in dealing with inconsistent temporal scale spectrum and relevant features.

(iii) Context selective long skip connections. We transmit only the microphone spectrum information, unlike a CRN-based AEC backbone network that sends all encoded results to the decoder. We believe that introducing historical frame information during the decoding stage can interfere with the masking estimation.

*Equal contribution.

10.21437/Interspeech.2024-1355

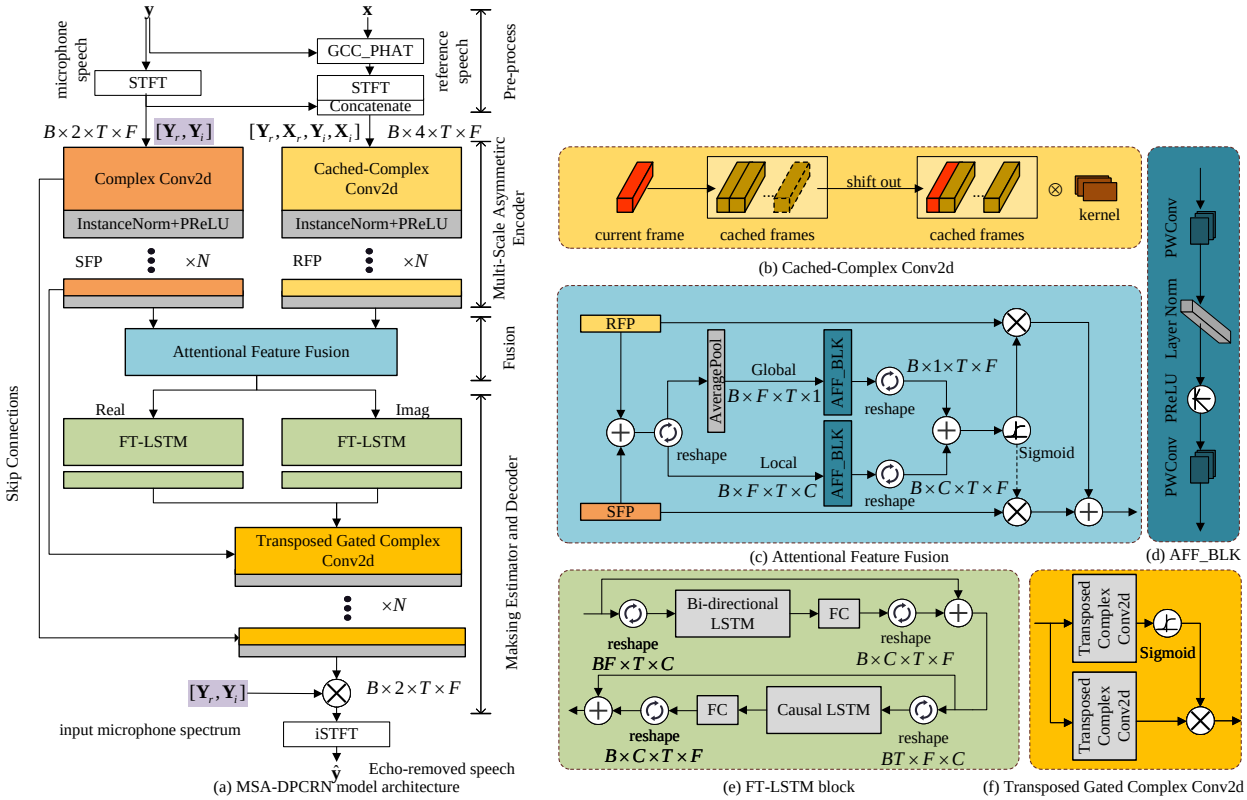

Figure 1: *(a) Overall architecture of our proposed MSA-DPCRN model. (b)-(f) Details of each block used in MSA-DPCRN.*

2. Problem formulation

In the AEC system, the far-end reference signal $x(n)$ is played by a loudspeaker, involving nonlinear distortions. The echo signal $d(n) = x(n - \Delta) * h(n)$ is generated when the microphone picks the delayed reference signal where the delay Δ is caused by the transmission path denoted as h. Besides the echo signal, the microphone also picks the near-end speech $s(n)$ and noise $v(n)$; therefore, microphone signal $y(n)$ can be expressed as:

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

The goal of the AEC system is to remove the echo $d(n)$ and noise $v(n)$ from the microphone signal $y(n)$ without distorting the target, obtaining the clean near-end speech $s(n)$.

3. Proposed method

In this section, we will describe the details of the proposed methods. The input reference and microphone signals are first aligned using the Generalized Cross-Correlation with Phase Transform (GCC-PHAT) algorithm [20]. Figure 1 illustrates the overall architecture of the MSA-DPCRN model which consists of a multi-scale asymmetric encoder, a feature fusion block, and masking estimating and applying blocks.

3.1. Multi-scale asymmetric encoder

To alleviate the feature encoding problems arising from the consistent temporal scales across different input features, we propose a multi-scale asymmetric dual-path encoder, a simple yet effective scheme to extract features across different fields.

Assuming at time t, the SFP takes the microphone spectrum feature $\mathbf{Y}(t) = [Y(t,k)] \in \mathbb{R}^{2\times 1\times F}, k \in [0,\dots,F-1]$ as input, where F is the number of the frequency dimension after STFT, primarily focusing on the spectrum context within the current frame. The RFP is employed to extract relevant features within the last L frames between the microphone $[\mathbf{Y}(t-L+1),\dots,\mathbf{Y}(t)] \in \mathbb{R}^{2\times L\times F}$ and reference signals $[\mathbf{X}(t-L+1),\dots,\mathbf{X}(t)] \in \mathbb{R}^{2\times L\times F}$ by cached convolution. The working flow of a frame-to-frame cached convolution layer is shown in Figure 1(b). This layer maintains a buffer to cache the features of the last L frames. Both the SFP and RFP encoding paths are 4-layer stacked complex convolution blocks [21] and the first block can be expressed as:

$$\begin{aligned}\mathbf{R}_{i=0}(t) &= \delta\left(\mathcal{B}\left(\begin{bmatrix}\mathbf{X}(t-L+1),\dots,\mathbf{X}(t)\\ \mathbf{Y}(t-L+1),\dots,\mathbf{Y}(t)\end{bmatrix} * \Phi_i\right)\right),\\ \mathbf{S}_{i=0}(t) &= \delta\left(\mathcal{B}\left(\mathbf{Y}(t) * \Theta_i\right)\right),\end{aligned} \tag{2}$$

where $\mathbf{S}_i(t), \mathbf{R}_i(t) \in \mathbb{R}^{C\times 1\times F'}$ indicates the i-th layer output of SFP and RFP encoding paths, respectively. Except for the first layer $i = 0$, the subsequent layer takes the output of the previous layer as its input. The δ denotes the PReLU activation and the $\mathcal{B}$ is the instance normalization that normalizes along the channel and frequency dimensions. The $*$ denotes the convolution operation, and the kernel of i-th layer is denoted as $\Theta_i \in \mathbb{R}^{1\times m}, \Phi_i \in \mathbb{R}^{L\times l}$, in which the m, l refer to the filter width along the frequency dimension, for SFP and RFP encoding path, respectively. By setting the kernel height of SFP and RFP to 1 and L, we can achieve intra-frame convolution and inter-frame convolution, respectively.

From the design, it can be observed that the intra-frame convolution lets the output features of the SFP contain more fine spectrum details. On the contrary, the inter-frame cached convolution filters spectrum features across multiple frames, containing more global features in RFP output. As the value of L increases, the RFP includes more historical frames, leading to a higher presence of global feature components and a lower resolution of local details.

3.2. Attentional feature fusion

To fuse features with varying scales and distinct spectrum contexts and maintain the original dynamic range of the features, we propose a frequency-wise AFF block of which the architecture is shown in Figures 1(c) and (d).

Given T frames spectrum features of SFP and RFP denoted as $\mathbf{S}, \mathbf{R} \in \mathbb{R}^{C\times T\times F'}$ with C channels and F' down-sampled frequency bins, we first integrated and reshaped the two features using the EAF method, and obtained the intermediate feature represented as $\mathbf{Z} = \text{reshape}(\mathbf{S}+\mathbf{R}) \in \mathbb{R}^{F'\times T\times C}$. The intermediate feature will be sent into two point-wise convolution (PWConv) blocks, denoted as global context extractor $\mathcal{M}(\cdot;\theta)$ and local context extractor $\mathcal{M}(\cdot;\phi)$, to learn the fusion weight that combines the multi-scale features. The PWConv blocks can be computed as:

$$\mathcal{M}(\cdot;\theta/\phi) = \text{PWConv}(\delta(\mathcal{B}'(\text{PWConv}(\cdot;\theta_1/\phi_1)));\theta_2/\phi_2), \quad (3)$$

where the δ and $\mathcal{B}'$ represent the PReLU activation and Layer Normalization, respectively. The global context extractor takes the average frequency features across all channels, also known as sub-bands, as input and condenses the context of each frequency into a scalar value. This scale value is a coarse representation of features across different scales and sub-bands. We utilize the local context extractor to extract detailed representations for every frequency component of each channel. At last, the attention weights $\mathbf{W} \in \mathbb{R}^{C\times T\times F'}$ for each frequency component are computed as:

$$\mathbf{W} = \text{reshape}(\sigma(\mathcal{M}(\mathcal{G}(\mathbf{Z});\theta) \oplus \mathcal{M}(\mathbf{Z};\phi))), \quad (4)$$

where $\mathcal{G}(\mathbf{Z}) = \frac{1}{C}\sum_{i=1}^{C}\mathbf{Z}[...,i] \in \mathbb{R}^{F'\times T\times 1}$ denotes the average pool operation across the channel dimension for each frequency. The $\oplus$ means the broadcasting addition. With the soft attention weights, the refined features $\mathbf{Z}' \in \mathbb{R}^{C\times T\times F'}$ are selected from the input features of SFP $\mathbf{S}$ and RFP $\mathbf{R}$:

$$\mathbf{Z}' = \mathbf{W}\odot\mathbf{R} + (\mathbf{I}-\mathbf{W})\odot\mathbf{S}, \quad (5)$$

where $\odot$ denotes an element-wise multiply.

3.3. Masking estimating and applying blocks

In our model, the real and imaginary parts of fused features are fed into the dual-stacked FT-LSTM [22] blocks, respectively, followed by an encoder to estimate the masking. As Figure 1(e) shows, the FT-LSTM block is composed of a bi-directional LSTM (F-LSTM) followed by a causal LSTM (T-LSTM). The decoders in the MSA-DPCRN model are built upon the transpose gated conv2d block shown in Figure 1(f). We apply the estimated masking $M_{r,i}$ to the input microphone spectrum $Y_{r,i}$ using the DCCRN-C [21] methods:

$$\hat{Y} = (Y_r * M_r - Y_i * M_i) + j(Y_r * M_r + Y_i * M_i), \quad (6)$$

where the subscripts r, i represent the real and imaginary parts of the component, respectively. $\hat{Y}$ is the estimated spectrum of near-end speech.

3.4. Training targets and loss function

MSA-DPCRN is a masking-based model as many previous studies [23, 24] demonstrate that masking-based models outperform the mapping-base models. We introduced a multi-resolution STFT loss [25] over the spectrogram magnitudes to optimize the MSA-DPCRN model. Given the clean near-end speech $\mathbf{y}$ and the estimated speech $\hat{\mathbf{y}}$ of the MSA-DPCRN model, an STFT loss can be computed as:

$$\begin{aligned}
\mathcal{L}_{stft}(\mathbf{y},\hat{\mathbf{y}}) &= 0.5\times(\mathcal{L}_{sc}(\mathbf{y},\hat{\mathbf{y}}) + \mathcal{L}_{mag}(\mathbf{y},\hat{\mathbf{y}})),\\
\mathcal{L}_{sc}(\mathbf{y},\hat{\mathbf{y}}) &= \frac{\||\text{STFT}(\mathbf{y})| - |\text{STFT}(\hat{\mathbf{y}})|\|_F}{\||\text{STFT}(\mathbf{y})|\|_F},\\
\mathcal{L}_{mag}(\mathbf{y},\hat{\mathbf{y}}) &= \frac{1}{T}\|\log|\text{STFT}(\mathbf{y})| - \log|\text{STFT}(\hat{\mathbf{y}})|\|_1, \quad (7)
\end{aligned}$$

where the $\|\cdot\|_F$ and $\|\cdot\|_1$ represent the Frobenius and L_1 norms, respectively. The final multi-resolution loss function is an average of three STFT losses with frame sizes $\in \{1024, 512, 256\}$ and corresponding hop lengths $\in \{512, 256, 128\}$.

4. Experiments

4.1. Datasets

We utilize the AEC Challenge [26] and the DNS Challenge [27] wide-band datasets to construct our training dataset. To obtain ground truth labels, we only utilize the far-end single-talk (FE) samples from the AEC Challenge including 20 h of real recordings and 27 h of synthesized samples. The DNS Challenge dataset consists of 544 h of speech samples and 181 h of noise samples. We generated a training dataset comprising 100 h of double-talk (DT) and 30 h of near-end single-talk (NE) scenarios. Each sample in the dataset has a duration of 10 s.

We generated noisy speech with a probability of 0.3 when creating the NE samples. The signal-to-noise ratio (SNR) is chosen from a uniform distribution ranging from 10 dB to 30 dB. To ensure the effectiveness, we discard the samples with an activation ratio lower than 0.6 or an average energy below -30 dB. To generate the DT samples, we randomly pick reference and corresponding echo signals from the AEC Challenge dataset. Then, synthesized the selected echo with the NE speech using a randomly selected signal-to-echo ratio (SER) from a uniform distribution ranging from -15 dB to 18 dB.

To test the performance and generalization ability of models, we conducted experiments on three datasets, including two blind test datasets provided by the AEC Challenge, as well as a synthetic dataset recorded and synthesized by ourselves. We randomly select samples from the test-clean subset of the LibriSpeech corpus [28] and play these audios to capture the corresponding echo signals under the FE scenario. Next, we synthesized DT test samples with specific SERs ranging from -10 dB to 10 dB with an increment of 5 dB. Each SER includes 1 h of speech. The model inference code, scripts used to synthesize samples, and our recorded samples will be made publicly on https://github.com/deepnetni/msa-dpcrn.

4.2. Compared models and ablation study

We consider: AEC Challenge wide-band baseline model [26], which concatenates the log power spectrum of the reference and microphone signal as inputs. Next, DTLN-AEC [4], a model using two symmetric encoding dual paths. Then, -w-skip-rel, -w-fusion-EAF, and -w/o-SFP variants. The -w-skip-rel model is similar to MSA-DPCRN but builds skip connections between

Table 1: *Subjective ratings of ITU-T P.831 on the two wide-band blind real test sets, ICASSP* and Interspeech†, of the AEC Challenge computed by AECMOS. Other-MOS means more ratings related to other degradations. The best score of each item is shown in boldface. All the models are real-time causal models.* ‡ *denotes the model we proposed.*

Models	# Params / Units	Flops	Echo-MOS*		Other-MOS*		Echo-MOS†			Other-MOS†		
			clean	noisy	clean	noisy	DT	FE	NE	DT	FE	NE
Baseline [26]	1.3 M	—	4.569	4.336	**4.435**	4.287	4.698	4.337	4.997	**2.4**	4.999	**4.199**
DTLN-AEC [4]	1.8 M	—	4.44	4.265	4.246	4.189	4.597	4.008	4.998	2.058	4.999	4.006
DPCRN-refine	4.243 M	4.571 G	4.616	4.471	4.334	4.226	4.687	4.279	4.998	1.7	4.999	4.061
MSA-DPCRN‡	1.048 M	4.585 G	**4.631**	**4.511**	4.401	4.312	**4.738**	**4.388**	**4.998**	1.823	4.999	3.928
-w-skip-rel	1.048 M	4.585 G	4.616	4.481	4.397	4.238	4.685	4.316	4.998	1.883	4.999	4.075
-w-fusion-EAF	1.03 M	4.519 G	4.624	4.505	4.409	**4.361**	4.711	4.312	4.997	2.03	4.999	4.181
-w/o-SFP	0.97 M	4.312 G	4.625	4.497	4.358	4.291	4.731	4.311	4.998	2.004	4.999	4.061

Table 2: *Ablation on the simulated test set with PESQ, STOI, and SI-SNR objective metrics at different SERs.*

Models	Metric	-10 dB	-5 dB	0 dB	5 dB	10 dB
Unprocessed	PESQ	1.115	1.152	1.244	1.389	1.625
MSA-DPCRN‡		**2.116**	**2.569**	**3.018**	**3.387**	**3.702**
-w-skip-rel		2.028	2.449	2.897	3.274	3.617
-w-fusion-EAF		2.067	2.512	2.966	3.332	3.655
-w/o-SFP		1.9	2.351	2.83	3.236	3.589
Unprocessed	STOI	0.576	0.667	0.760	0.838	0.896
MSA-DPCRN‡		**0.906**	**0.936**	**0.959**	**0.972**	**0.978**
-w-skip-rel		0.899	0.931	0.956	0.971	0.977
-w-fusion-EAF		0.903	0.934	0.958	0.972	0.978
-w/o-SFP		0.888	0.925	0.953	0.969	0.976
Unprocessed	SI-SNR	-9.986	-4.997	0.008	5.011	10.009
MSA-DPCRN‡		**8.434**	**10.923**	**13.777**	**16.589**	**19.493**
-w-skip-rel		8.124	10.597	13.489	16.334	19.3
-w-fusion-EAF		8.4	10.849	13.743	16.528	19.442
-w/o-SFP		8.078	10.627	13.562	16.398	19.365

the RFP and decoder layers. The -w-fusion-EAF model replaces the AFF fusion block in MSA-DPCRN with EAF methods. The -w/o-SFP model eliminates the SFP path and introduces skip connections between the RFP and decoder layers. Finally, DPCRN-refine, employing the refinement fusion [10] to -w/o-SFP model. We utilize the AECMOS [29] model to compute the metrics since blind test sets don't contain a ground truth sample. For the simulated test set, we select perceptual evaluation of speech quality (PESQ) [30], short-time objective intelligibility (STOI) [31], and scale-invariant SNR (SI-SNR) [32] as objective metrics.

4.3. Parameters setup

The frame and hop length are set to 32 ms and 16 ms, respectively. We employ the Adam optimizer with a learning rate of $5e^{-4}$ decayed by 0.8 every 20 epochs. The kernel size of the stacked four-layer complex convolutions in SFP and RFP is set to $(1, 5)$ and $(3, 5)$, respectively, with channels $\in \{16, 32, 64, 128\}$ and corresponding strides $\in \{2, 2, 1, 1\}$. The hidden size of the FT-LSTM is set to 128. The MSA-DPCRN and its ablation variants are trained for 30 epochs using a batch size of 4.

5. Experimental results and discussion

As demonstrated in Table 1, our proposed model, MSA-DPCRN, outperforms all compared methods in terms of the Echo-MOS metric, and our pre-RNN fusion AFF outperforms the post-RNN fusion refinement method compared with the DPCRN-refine model. The average Echo-MOS in the ICASSP test set increased from 4.453 to 4.571, and in the Interspeech test set, it improved from 4.677 to 4.708 compared to the baseline model. However, the models we trained did not perform as well as the baseline model when considering the Other-MOS metric related to degradations about noise, distortions, cut-outs, etc. The Other-MOS decreased from 4.361 and 3.866 to 4.357 and 3.583, respectively. This discrepancy could be attributed to the mismatch data distributions between the training dataset and the Interspeech test set, as all models with various architectures that we trained perform worse on the Interspeech test set compared to the baseline model; however, their performance on the ICASSP test set is extremely close with only 0.04 decrement. Additionally, our synthesized script does not specifically focus on these tasks. We generated noisy data with a probability of 0.3 and SNR$\in [10, 30]$ dB, meaning the majority of the training dataset consists of clean speech, and even the noisy samples are not in adverse conditions.

The ablation study results are shown in Table 2. The performance metrics show an increasing trend as SER increases from -10 dB to 10 dB. Moreover, the performance of the MSA-DPCRN model $>$ -w-fusion-EAF $>$ -w-skip-rel $>$ -w/o-SFP across all the evaluated metrics, which is consistent with our expectations. Specifically, the -w-fusion-EAF model outperforms the -w/o-SFP model with an average increment of 0.125 in PESQ, 0.007 in STOI, and 0.186 in SI-SNR, which demonstrates that using an asymmetric dual encoding path is beneficial for improving the model's performance. The AFF method further improves the performance in all metrics when comparing the MSA-DPCRN and -w-fusion-EAF models, particularly in terms of PESQ. The average PESQ improved by 1.79% increasing from 2.906 to 2.958. Additionally, the MSA-DPCRN outperforms the -w-skip-rel model indicating that passing the microphone spectrum feature through skip connections is more effective than relevant features. The inference time of each frame of the proposed MSA-DPCRN model is 2 ms using the i9-13900KF processor.

6. Conclusion

In this study, we proposed an MSA-DPCRN AEC model. This model utilizes two distinct encoding paths, SFP and RFP, to extract spectrum and relevant features that have varying temporal scales considering their characteristic properties. Additionally, we proposed a frequency-wise AFF block to merge the two features while preserving the dynamic range. The experiments show that both two methods improve the model's capability to eliminate echo, and our model outperforms the AEC Challenge baseline model in terms of the Echo-MOS. In future work, we will explore whether our method is suitable for inputting other features and explore more cascade schemes.

7. References

[1] T. O'Malley, A. Narayanan, Q. Wang, A. Park, J. Walker, and N. Howard, "A conformer-based asr frontend for joint acoustic echo cancellation, speech enhancement and speech separation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 304–311.

[2] Z. Lv, F. Poiesi, Q. Dong, J. Lloret, and H. Song, "Deep learning for intelligent human–computer interaction," *Applied Sciences*, vol. 12, no. 22, p. 11457, 2022.

[3] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Interspeech*, vol. 2018, 2018, pp. 3229–3233.

[4] N. L. Westhausen and B. T. Meyer, "Acoustic echo cancellation with the dual-signal transformation lstm network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7138–7142.

[5] L. Ma, S. Yang, Y. Gong, X. Wang, and Z. Wu, "Echofilter: End-to-end neural network for acoustic echo cancellation," *arXiv preprint arXiv:2105.14666*, 2021.

[6] H. Zhang, K. Tan, and D. Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions." in *Interspeech*, 2019, pp. 4255–4259.

[7] C. Zhang and X. Zhang, "A robust and cascaded acoustic echo cancellation based on deep learning." in *INTERSPEECH*, 2020, pp. 3940–3944.

[8] S. Zhang, Y. Kong, S. Lv, Y. Hu, and L. Xie, "Ft-lstm based complex network for joint acoustic echo cancellation and speech enhancement," *arXiv preprint arXiv:2106.07577*, 2021.

[9] R. Peng, L. Cheng, C. Zheng, and X. Li, "Acoustic echo cancellation using deep complex neural network with nonlinear magnitude compression and phase information." in *Interspeech*, 2021, pp. 4768–4772.

[10] F. Cui, L. Guo, W. Li, P. Gao, and Y. Wang, "Multi-scale refinement network based acoustic echo cancellation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9132–9136.

[11] A. Fazel, M. El-Khamy, and J. Lee, "Cad-aec: Context-aware deep acoustic echo cancellation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6919–6923.

[12] G. Zhang, L. Yu, C. Wang, and J. 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)*. IEEE, 2022, pp. 9122–9126.

[13] S. Zhang, Z. Wang, J. Sun, Y. Fu, B. Tian, Q. Fu, and L. Xie, "Multi-task deep residual echo suppression with echo-aware loss," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9127–9131.

[14] J. Zhou, Y. Gao, and S. Zhang, "Ultra-low complexity residue echo and noise suppression based on recurrent neural network," in *National Conference on Man-Machine Speech Communication*. Springer, 2023, pp. 1–8.

[15] B. S. Nollett and D. L. Jones, "Nonlinear echo cancellation for hands-free speakerphones," *Proc. NSIP'97*, pp. 8–10, 1997.

[16] A. Stenger and W. Kellermann, "Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling," *Signal Processing*, vol. 80, no. 9, pp. 1747–1760, 2000.

[17] E. Indenbom, N. Ristea, A. Saabas, T. Pärnamaa, and J. Gužvin, "Deep model with built-in cross-attention alignment for acoustic echo cancellation," 2023.

[18] Y. Liu, Y. Shi, Y. Li, K. Kalgaonkar, S. Srinivasan, and X. Lei, "Sca: Streaming cross-attention alignment for echo cancellation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[19] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3560–3569.

[20] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[21] 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.

[22] X. Le, H. Chen, K. Chen, and J. Lu, "Dpcrn: Dual-path convolution recurrent network for single channel speech enhancement," *arXiv preprint arXiv:2107.05429*, 2021.

[23] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[24] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[25] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[26] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, H. Gamper, S. Braun, K. Sorensen, and R. Aichner, "Icassp 2022 acoustic echo cancellation challenge," in *ICASSP 2022*, 2022.

[27] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *INTERSPEECH*, 2020.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[29] M. Purin, S. Sootla, M. Sponza, A. Saabas, and R. Cutler, "Aecmos: A speech quality assessment metric for echo impairment," *arXiv preprint arXiv:2110.03010*, 2021.

[30] 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," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[31] 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.

[32] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.