MULTI-SCALE REFINEMENT NETWORK BASED ACOUSTIC ECHO CANCELLATION

Fan Cui, Liyong Guo, Wenfeng Li, Peng Gao, Yujun Wang

Xiaomi Inc., Beijing, China

ABSTRACT

Recently, deep encoder-decoder networks have shown outstanding performance in acoustic echo cancellation (AEC). However, the subsampling operations like convolution striding in the encoder layers significantly decrease the feature resolution lead to fine-grained information loss. This paper proposes an encoder-decoder network for acoustic echo cancellation with mutli-scale refinement paths to exploit the information at different feature scales. In the encoder stage, high-level features are obtained to get a coarse result. Then, the decoder layers with multiple refinement paths can directly refine the result with fine-grained features. Refinement paths with different feature scales are combined by learnable weights. The experimental results show that using the proposed multi-scale refinement structure can significantly improve the objective criteria. In the ICASSP 2022 Acoustic echo cancellation Challenge, our submitted system achieves an overall MOS score of 4.439 with 4.37 million parameters at a system latency of 40ms.

Index Terms— acoustic echo cancellation, encoder-decoder, multi-scale

1. INTRODUCTION

Acoustic echo refers to the phenomenon that occurs when a microphone picks up the far-end signal that is played by a loudspeaker. This phenomenon can cause a slight annoyance or a significant breakdown in a communication system.

Acoustic echo cancellation is designed to remove echo, reverberation and environment noise. The traditional AEC approach is based on adaptive filters to model the echo impulse response [1, 2]. Based on the predicted echo impulse response, the estimated echo can be subtracted from the microphone signal to obtain the near-end clean speech. The normalized least squared (NLMS) algorithms [3] are the most popular solutions. However, those methods model linear echo and cannot process non-linear echo path distortions and environment noise picked up by the microphone, which may degrade both speech intelligibility and speech quality. Especially, a double-talk detector is need to stop adapting the filter during both near-end and far-end are talking simultaneously [4]. The Near-end speech distortion and residual echo suppression is a difficult tradeoff for a traditional AEC system.

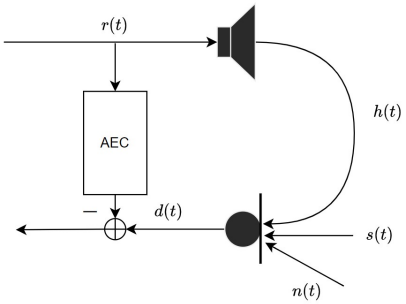

Fig. 1. Diagram of the acoustic echo cancellation system.

Recently, with the development of deep learning, AEC is formulated as a supervised learning problem that estimate the ideal ratio mask from near-end and far-end signals for removing non-linear residual echo and noise [5, 6, 7, 8]. Although deep learning based AEC method has been intensively investigated, it is still difficult to implement in real applications. Because most methods are noncausal, computational demanding, and not robust.

We propose a new model structure for real-time AEC, where the loudspeaker and microphone signals are combined in encoder layers to predict a coarse ideal ratio mask. The decoder layers combine the encoder features with short connections. Refinement paths are designed to obtain a residual ideal ratio mask for every layer with learnable weights to exploit useful information from different feature scales.

The remainder of this paper is organized as follows: Section 2 introduces the formulates the problem. In Section 3, the framework of the proposed method is shown. The experimental setup is presented in Section 4. The experimental results and analyses are shown in Section 5. Section 6 is the conclusion part.

2. PROBLEM FORMULATION

Figure 1 shows the diagram of traditional AEC system. The far-end microphone signal $r(t)$, which is played by the loudspeaker with echo path $h(t)$, can be picked up by a microphone. The near-end signal $d(t)$ is recorded by the micro-

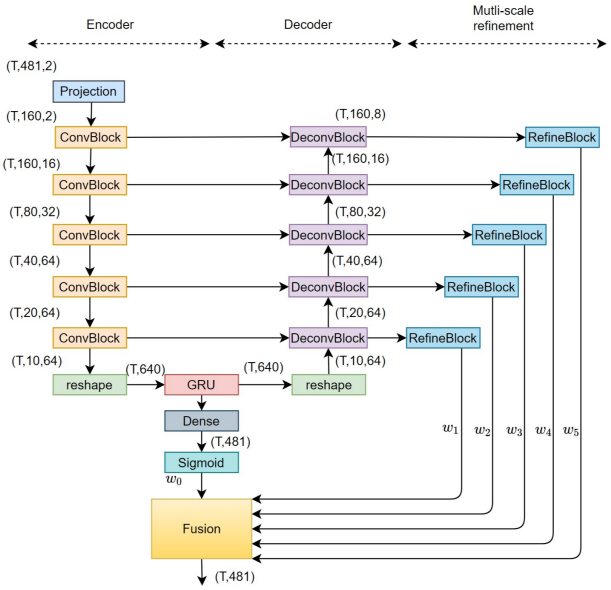

Fig. 2. The proposed model structure.

phone can be represented as :

$$d(t) = r(t - \triangle t) * h(t) + n(t) + s(t) \quad (1)$$

where $*$ denotes convolution, t represents the discrete-time index, $s(t)$ is the near-end clean speech, $n(t)$ is the additive environment noise picked up by the microphone. The far-end reference signal is delayed by a time $\triangle t$, which is depending on the latency of the device. The proposed AEC method is completely performed by an encoder-decoder deep neural network. The input signals are the far-end reference signal $r(t)$ and the near-end microphone signal $d(t)$, which is similar to the traditional method. However, rather than predicting the echo path, the deep network predicts the magnitude spectrum of the near-end speech $s(t)$ directly.

3. PROPOSED METHOD

Due to the promising performance of convolutional encoder-decoder structure [9, 10] for audio signal processing. We adopt it in our model structure. Figure 2 shows the proposed model structure. It mainly consists of three parts, namely the encoder module, decoder module, and multi-scale refinement module. Each module is composed of five interconnected blocks. The whole method is operated in the magnitude domain. Firstly, we decouple the complex spectrum into magnitude and phase, the magnitude spectrum of the far-end reference signal and near-end signal are concatenated as the input $X \in R^{T \times F \times C}$, where T represents time domain, $F = 481$ represents frequency domain with window size 20 ms for sample rate 48kHz, and $C = 2$ represents far-end and near-end signal, leaving the phase unchanged. To

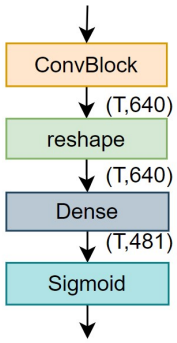

Fig. 3. The structure of refinement block.

reduce computational complexity, a linear projection layer is adopted to reduce the frequency dimension from 481 to 160. The encoder module is composed by five convolutional blocks followed by a GRU layer. Each convolutional block contains a convolution layer followed by batch normalization and RELU. The kernel size of the five encoder blocks are $(3, 3), (1, 5), (5, 1), (1, 5), (5, 1)$, and the stride of the encoder blocks are $(1, 1), (1, 2), (1, 2), (1, 2), (1, 2)$. For real-time inference, only the first encoder convolutional layer look ahead one frame, the other four layers are implemented by casual convolutional layer [11]. A GRU block is used to capture long-term information. After the GRU block, A dense layer with sigmoid activation is applied to get a coarse ideal ratio mask.

In the decoder part, each decoder layer is composed by deconvolutional layer followed by batch normalization and RELU. Every decoder layer is connected by skip connection with corresponding encoder layer, which is utilized to mitigate the information loss. The model configure of the decoder part is similar to the encoder part, where the convolutional layer is replaced by the deconvolutional layer to recover the original feature size. The convolutional striding operations can decrease feature resolution, which reduce the computational complexity and increase context field. However, decreasing the feature resolution leads to fine-grained information loss. Here, a refinement block is adopted to get a residual ideal ratio mask from different feature scale for each decoder layer. Figure 3 shows the diagram of the refinement block, the refinement block is composed by a convolutional block with kernel size $(3, 3)$, and a reshape layer is adopted to reshape the feature map, which can be projected to residual ideal ratio mask by a dense and sigmoid layer. Then, the coarse ideal ratio mask is combined with the residual ideal ratio masks by a set of learnable weights $w_0, w_1, w_2, w_3, w_4, w_5$. Finally, the ideal ratio mask is applied to the original magnitude spectrum of near-end signal. By combining the unchanged phase, the predicted near-end speech is obtained. For training stability, a chained optimization strategy is applied by separating the encoder, decoder and refinement module optimization.

4. EXPERIMENTS SETUP

4.1. Datasets and Data Augmentation

The ICASSP 2022 AEC challenge provides recordings from more than 7,500 real audio devices and human speakers in real environments [12]. It covers both single talk and double-talk scenarios with various nonlinear distortions. The default sample rate is 48 kHz. In the training stage, real recordings and synthetic samples are adopted. To have a ground truth for training, we only use far-end single-talk real recordings. Additionally, the DNS challenge [13] noise dataset is used to increase the variance of noise. To improve the robustness of the proposed method, we generate training sample pairs by randomly selecting the far-end reference signal, the near-end speech signal, the echo path and the additive noise independently. For each pair, the signal-to-echo ratio (SER) and signal-to-noise ratio (SNR) between the echo, the desired signal and the environment noise are selected from a uniform distribution between -10dB to 10dB. To stimulate various transmission effects, a delay variance is randomly chosen from a uniform distribution between -100ms to 10ms.

To compare the performance of the proposed method, we generate a small test set (2000 samples), contains far-end single-talk, near-end single-talk and double-talk scenarios. The far-end reference signal and the echo signal pair are selected from ICASSP 2022 AEC challenge far-end single talk test set. The near-end signals are separated from the provided synthesis dataset. The noises are selected from MUSAN [14] and RIRs are generated using the image method [15]. For the double-talk test set, SER is randomly selected from $-10dB, -5dB, 0dB, 5dB, 10dB$. For the noisy scenarios, SNR is randomly selected from $-5dB, 0dB, 5dB, 10dB, 15dB$.

4.2. Training Setup

The near-end microphone signal and far-end reference signal are transformed into frequency domain by applying Hanning window short time Fourier transform (STFT), the window size is 20 ms and the window shift is 10 ms. The delay of the proposed method is 40 ms, because it needs to look ahead one frame.

The proposed model is trained to optimize the loss function with the Adam optimizer. The loss function has two targets. The first part is MSE loss [16] applied to evaluate the magnitude spectrum similarity between predicted spectrum and the ground truth. The second part is SI-SNR [17] used to evaluate the time domain processing result. The learning rate is 0.0003. Weight decay and gradient norm are used to avoid overfitting. The whole framework takes a chained optimization strategy and can be divided into two parts. In the first part, the optimization is implemented to get a coarse ideal ratio mask from the encoder part. In the second stage, the decoder and refinement module are combined to refine the final ideal ratio mask. The number of trainable parameters of the proposed model is 4.27M. 10 epochs are set for the first training stage, 30 epochs for the second stage, with a batch size of 16. The best model has the minimum validation loss.

4.3. Objective and Subjective Audio Quality Evaluation

Traditional objective metrics such as the perceptual evaluation of speech quality (PESQ)[18], echo return loss enhancement (ERLE)[19] are often used for evaluating speech quality. However, those methods do not correlate well with subjective speech quality for reverberation, additive noise and non-linear distortions. The ITU P.831 crowd-sourcing framework[20] is proposed on the Amazon Mechanical Turk platform for subjective evaluation. There are four scenarios: single-talk near-end, single-talk far-end, double-talk echo and double-talk other disturbances.

5. RESULTS AND DISCUSSION

5.1. Compared Methods

We compare the proposed several frameworks with the baseline system. Different model configurations are shown as followed:

- baseline: the baseline model is provided by ICASSP 2022 AEC challenge [12], which is a recurrent neural network with gated recurrent units takes concatenated log power spectral features of the microphone signal and far end signal as input, and outputs a spectral suppression mask.
- Encoder: this model has only the encoder part shown in Figure 2 without the decoder part and the refinement structures. The combination weights $w_0, w_1, w_2, w_3, w_4, w_5$ are 1.0, 0.0, 0.0, 0.0, 0.0, 0.0.
- ED: it is the encoder-decoder model. Compared with Encoder model, it contains decoder part. The combination weights $w_0, w_1, w_2, w_3, w_4, w_5$ are 0.0, 0.0, 0.0, 0.0, 0.0, 1.0.
- EDAMR: the model is proposed with refinement structure. The refinement results are combined by averaging. The combination weights $w_0, w_1, w_2, w_3, w_4, w_5$ are 1/6, 1/6, 1/6, 1/6, 1/6, 1/6.
- EDLMR: this is our proposed model with learnable combination weights.

5.2. Result Analysis

For comparing the performance of the baseline system with the proposed Encoder, ED, EDAMR, EDLMR models, the

Table 1. Subjective ratings of ITU-T P.831 on the blind real testset.

	ST NE MOS	ST FE Echo DMOS	DT Echo DMOS	DT Other DMOS
Baseline	4.152	4.563	4.122	3.563
Ours	4.317	4.759	4.654	4.025

experiments are developed on three scenarios which covers near-end single-talk, far-end single-talk and double-talk cases.

In Table 1, the evaluation results of ITU-T P.831 [20] on the ICASSP 2022 AEC Challenge blind test set are shown. The proposed method achieves overall DMOS 4.439. Note that our model shows an averaging 0.339 MOS points improvement over the baseline in the subjective ratings.

In Table 2, we compare our proposed frameworks with baseline system in near-end single-talk scenario, which can be influenced by the residual noise and speech distortion. The results show that even the Encoder model can get a higher PESQ score than the baseline system. By adding the decoder part in the ED model, the result has a slight improvement. In the proposed EDAMR and EDLMR system, we combine the multiple refinement paths, which lead to a great gain for PESQ score. Especially, when the refinement paths are combined by learnable weights, the PESQ score can be further improved. The results show that our method outperforms baseline by a large margin.

Double-talk scenario is the most difficult case for AEC. For verifying the performance of the proposed method in double-talk scenario, we develop two experiments. In Table 3, different models are compared with different SER (-10dB, -5dB, 0dB, 5dB, 10dB). Table 4 shows the results of different methods with different SER in noisy situation, the SNR is randomly selected from a uniform distribution between -5dB to 5dB. The results show that the proposed EDLMR model achieves a considerable gain on PESQ. This suggests that our system is also robust for double-talk scenarios and has the potential to removing noise and echo from the mixture signal.

Table 5 shows the ERLE for the far-end single-talk scenario. For this scenario, it can be seen that the proposed EDLMR system greatly improves the ERLE, compared to the baseline system. These results indicate that the proposed method can preserve the speech quality while removing residual echo. Compared with ED model, the learnable combination of multi-scale refinement paths improve ERLE from 49.26dB to 52.79dB.

For this work, the window size is 20ms, the overlap is 10ms. The whole system needs to look ahead one frame. Therefore, the algorithm delay is 40ms, which satisfies the latency requirement for ICASSP 2022 AEC challenge. Additionally, the trainable parameters of the proposed model is 4.27 million.

Table 2. Objective results of near-end single-talk in terms of PESQ.

Metrics	PESQ				
SNR(dB)	-5	0	5	10	15
Noisy	1.127	1.259	1.483	1.812	2.192
Baseline	1.433	1.714	1.881	2.027	2.213
Encoder	1.645	1.921	2.209	2.420	2.560
ED	1.772	2.074	2.399	2.613	2.833
EDAMR	1.778	2.101	2.412	2.627	2.846
EDLMR	**1.825**	**2.114**	**2.443**	**2.638**	**2.852**

Table 3. Objective results of double-talk without noise in terms of PESQ.

Metrics	PESQ				
SER(dB)	-10	-5	0	5	10
Noisy	1.215	1.280	1.384	1.534	2.776
Baseline	1.415	1.731	1.792	2.110	2.245
Encoder	1.614	1.807	2.049	2.316	2.615
ED	1.801	2.015	2.250	2.494	2.751
EDAMR	1.811	2.034	2.276	2.503	2.778
EDLMR	**1.839**	**2.045**	**2.279**	**2.516**	**2.782**

Table 4. Objective results of double-talk with noise (SNR:-5dB 5dB) in terms of PESQ.

Metrics	PESQ (SNR:-5dB∼5dB)				
SER(dB)	-10	-5	0	5	10
Noisy	1.178	1.218	1.296	1.398	1.576
Baseline	1.399	1.637	1.781	1.993	2.197
Encoder	1.534	1.696	1.903	2.108	2.349
ED	1.727	1.909	2.118	2.311	2.513
EDAMR	1.750	1.930	2.142	2.324	2.532
EDLMR	**1.755**	**1.945**	**2.155**	**2.344**	**2.541**

Table 5. Objective results of far-end single-talk in terms of ERLE.

Metrics	Baseline	Encoder	ED	EDAMR	EDLMR
ERLE (dB)	45.13	49.26	51.37	51.34	**52.79**

6. CONCLUSIONS

In this paper, we propose a novel multi-scale refinement framework for acoustic echo cancellation task to suppress acoustic echo, reverberation and environmental noise. The proposed method combines refinement paths with learnable weights to exploit useful information from different feature scales. The experimental results show the effectiveness of the multi-scale refinement structure and the robustness of the proposed method in the P.831 evaluation.

7. REFERENCES

[1] Eberhard Hänsler and Gerhard Schmidt, *Acoustic echo and noise control: a practical approach*, John Wiley & Sons, 2005.

[2] Simon S Haykin, *Adaptive filter theory*, Pearson Education India, 2008.

[3] Kai Steinert, Martin Schonle, Christophe Beaugeant, and Tim Fingscheidt, "Hands-free system with low-delay subband acoustic echo control and noise reduction," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 1521–1524.

[4] 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.

[5] Hao Zhang and D Wang, "Deep learning for acoustic echo cancellation in noisy and double-talk scenarios," *Training*, vol. 161, no. 2, pp. 322, 2018.

[6] Renhua Peng, Linjuan Cheng, Chengshi Zheng, and Xiaodong Li, "Acoustic echo cancellation using deep complex neural network with nonlinear magnitude compression and phase information," *Proc. Interspeech 2021*, pp. 4768–4772, 2021.

[7] Amin Fazel, Mostafa El-Khamy, and Jungwon 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.

[8] Lu Ma, Hua Huang, Pei Zhao, and Tengrong Su, "Acoustic echo cancellation by combining adaptive digital filter and recurrent neural network," *arXiv preprint arXiv:2005.09237*, 2020.

[9] Andong Li, Wenzhe Liu, Xiaoxue Luo, Guochen Yu, Chengshi Zheng, and Xiaodong Li, "A simultaneous denoising and dereverberation framework with target decoupling," *Proc. Interspeech 2021*, pp. 2801–2805, 2021.

[10] Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[11] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.

[12] Ross Cutler, Ando Saabas, Tanel Parnamaa, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sørensen, and Robert Aichner, "Icassp 2022 acoustic echo cancellation challenge," .

[13] Chandan KA 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 *INTERSPEECH*, 2021.

[14] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[15] 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.

[16] Meet H Soni, Neil Shah, and Hemant A Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5039–5043.

[17] Chao Ma, Dongmei Li, and Xupeng Jia, "Optimal scale-invariant signal-to-noise ratio and curriculum learning for monaural multi-speaker speech separation in noisy environment," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 711–715.

[18] Antony W Rix, John G Beerends, Michael P Hollier, and Andries 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)*. IEEE, 2001, vol. 2, pp. 749–752.

[19] Christina Breining, P Dreiscitel, Eberhard Hansler, Andreas Mader, Bernhard Nitsch, Henning Puder, Thomas Schertler, Gerhard Schmidt, and Jan Tilp, "Acoustic echo control. an application of very-high-order adaptive filters," *IEEE signal processing Magazine*, vol. 16, no. 4, pp. 42–69, 1999.

[20] Ross Cutler, Babak Nadari, Markus Loide, Sten Sootla, and Ando Saabas, "Crowdsourcing approach for subjective evaluation of echo impairment," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 406–410.