A PROGRESSIVE NEURAL NETWORK FOR ACOUSTIC ECHO CANCELLATION

Zhuangqi Chen[1,2], *Xianjun Xia*[1], *Siyu Sun*[1,3], *Ziqian Wang*[4], *Cheng Chen*[1], *Guoliang Xie*[1]
Pingjian Zhang[2], *Yijian Xiao*[1]

[1]RTC Lab, ByteDance, China
[2]SSE, South China University of Technology, Guangzhou, China
[3]SEI, Wuhan University, Wuhan, China
[4]ASLP, Northwestern Polytechnical University, Xian, China

ABSTRACT

Acoustic echo cancellation is a key issue in hand-free communication systems. In this paper, we proposed a hybrid signal processing and deep echo cancellation method, where a two-stage neural network is designed to remove residual echo progressively. For the personalized acoustic echo cancellation, we proposed to decouple the tasks of echo cancellation and target speech extraction, and introduced a speaker attentive module for personalized separation, where the ECAPA-TDNN is used for speaker embedding generation. The proposed method (*ByteAudio-18*) ranked first on both Track 1 and Track 2 in ICASSP 2023 AEC Challenge.

Index Terms— acoustic echo cancellation, two-stage, personalized, speaker attentive module

1. INTRODUCTION

The ICASSP 2023 Grand Challenge - fourth Acoustic Echo Cancellation (AEC) Challenge [1] focuses on the tasks of the general and personalized AEC. It's more challenging to consider echo cancellation, noise suppression, and interfering speech removal simultaneously. In this paper, we introduce our system (*ByteAudio-18*) submitted to the Challenge. The submitted entry is a hybrid of linear AEC and deep echo suppression method. We propose a two stage neural network, which has been proven to be capable of reaching a higher upper bound of performance with less computation cost in our previous experiments. A CRN based light-weight module is first employed to remove partial echo, which helps to reduce the modeling burden of single stage model. The band-split RNN (BSRNN) [2] with a few tweaks is then applied to further remove residual echo and recover nearend speech. For the personalized AEC task, we propose to decouple the echo cancellation and target speech extraction. The speaker attentive module [3] is introduced to perform personalized separation in the BSRNN, where the speaker embedding extracted by the ECAPA-TDNN [4] is used. In addition, a data cleaning strategy is employed to filter the echo data to alleviate nearend speech distortion. According to the official result[1], the proposed general and personalized AEC method obtains a final score of 0.852 and 0.854 respectively, and ranked first on both Track 1 and Track 2.

2. PROPOSED METHOD

In a full-duplex communication system, the microphone signal can be formulated as $y(t) = s(t) + n(t) + F(x(t)) \in \mathbb{R}^{1\times S}$, where $s(t) \in \mathbb{R}^{1\times S}$ is the nearend speech signal which contains target speech $s_t(t)$ and potential interference speech $s_o(t)$, $n(t) \in \mathbb{R}^{1\times S}$ is the noise signal, $x(t) \in \mathbb{R}^{1\times S}$ is the reference signal, and $F(*)$ denotes the echo path. In this paper, the $s(t)$ is expected for the general AEC task, while only the $s_t(t)$ is desired for the personalized AEC task by introducing a target speaker embedding $e \in \mathbb{R}^{1\times N}$.

As depicted in Fig. 1, the proposed general AEC and personalized AEC methods share the same framework which consists of a preprocessing module for aligning the reference signal and preliminary linear echo removal, and a two-stage neural network for progressive target speech extraction.

2.1. Preprocessing

Given a reference signal $x(t)$ and a microphone signal $y(t)$, a preprocessing module is first applied to predict an aligned reference signal $x(t-\delta)$ and an error signal $d(t)$. The preprocessing module consists of a time delay compensation (TDC) block and a linear AEC (LAEC) block. TDC and LAEC blocks are performed on the subband features, which are generated by dividing the frequency features into K subbands. The TDC block is based on the subband cross-correlation, where several time delays are estimated in each subband and the final time delay δ is determined using a simple voting method. The LAEC is a subband adaptive filtering method based on the NLMS, which consists of two filters: pre-filter and post-filter. The post-filter is adaptive using a dynamic step size, and the pre-filter is a backup of the post-filer with stable status. The error signal $d(t)$ is finally selected by comparing the residual energy of the outputs of pre-filter and post-filer.

2.2. Two-stage NN for general AEC

Taking $d(t), x(t-\delta), y(t)$ as input, a two-stage neural network based residual echo suppressor (TSNNRES) is employed to predict the nearend speech signal $\hat{s}(t)$. To improve the awareness of the low energy nearend speech, the stacked compressed magnitude $I_{cprs} = stack([|D|^\alpha, |X|^\alpha, |Y|^\alpha]) \in \mathbb{R}^{3\times F\times T}$ is used as the network input, where $\alpha \in (0, 1]$ is the compression factor, and $D, X, Y \in \mathbb{C}^{F\times T}$ are the STFT representations of $d(t), x(t-\delta), y(t)$ respectively, where F, T are the number of frequency and time bins respectively.

The TSNNRES consists of two processing modules: a CRN based light-weight module for preliminary echo cancellation and noise suppression, and a BSRNN based post-processing module for better nearend speech signal reconstruction. The CRN based light-weight module consists of a band merging (BM) [5] block, a encoder, two dual-path GRU (DPGRU), a decoder, and a band spliting (BS) [5] block. More importantly, a voice activity detection (VAD) module is used for multi-task learning, which help to improve the awareness of nearend speech [6]. The CRN takes the I_{cprs} as input, and output a rough complex ideal ratio mask $\hat{M}_1 \in \mathbb{C}^{F\times T}$ of the error signal D and a nearend VAD probability $P_{vad} \in \mathbb{R}^{2\times T}$.

[1]https://www.microsoft.com/en-us/research/academic-program/acoustic-echo-cancellation-challenge-icassp-2023/results/

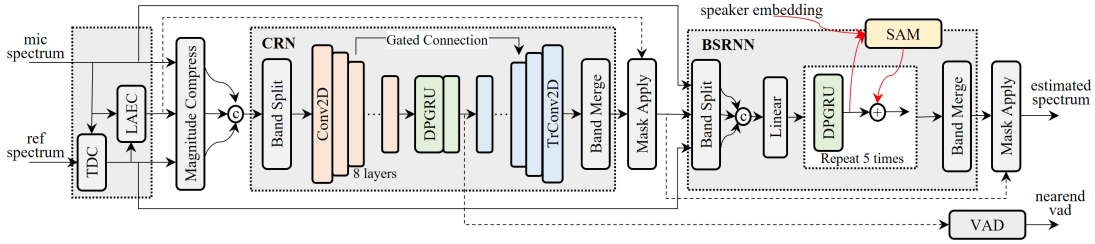

Fig. 1. The flowchart of proposed AEC, where SAM denotes the speaker attentive module only used for personalized separation.

Table 1. Comparison with other methods on the AEC Challenge blind test set, where 'P' denotes 'Personalized'.

Method	P	Subjective-MOS	WAcc	Final Score
baseline	N	4.013	0.649	0.736
Ours	N	4.433	0.823	0.852
Ours	Y	4.444	0.822	0.854

Since the BSRNN achieves state-of-the-art performance in speech enhancement [2], we adopt the BSRNN as the post-processing module with a few tweaks. The LSTM module is replaced by GRU and the hidden size is also reduced for computational efficiency.

2.3. Personalized AEC with ECAPA-TDNN

To perform the personalized AEC, we introduce the speaker attentive module (SAM) to make the TSNNRES predict the nearend target speech $\hat{s}_t(t)$. An additional speaker embedding e extracted by employing the ECAPA-TDNN is used. We propose to decouple the tasks of echo cancellation and target speaker extraction. To this end, the SAM is only performed at the BSRNN based post-processing stage, where the outputs of each hidden layer (i.e., the DPRNN layer) are applied. Given a speaker embedding $e \in \mathbb{R}^{B\times N}$ and a hidden layer output $h \in \mathbb{R}^{B\times C\times F\times T}$ where B, N, C are the batch size, the dimension of embedding and the number of channels, the SAM [3] outputs a personalized feature $h_p \in \mathbb{R}^{B\times C\times F\times T}$.

2.4. Loss function

The final loss to be optimized is

$$\mathcal{L}(S,\hat{S}) = \mathit{MAE}(|S|,|\hat{S}|) + \mathit{MAE}(S_r,\hat{S}_r) + \mathit{MAE}(S_i,\hat{S}_i) \quad (1)$$

$$loss_{AEC} = \omega\mathcal{L}(S,\hat{S}_1) + (1-\omega)\mathcal{L}(S,\hat{S}_2) \quad (2)$$

$$loss_{final} = loss_{AEC} + \beta CrossEntropy(P_{vad}, P) \quad (3)$$

where S denotes the groundtruth and $\hat{S}_1, \hat{S}_2$ denote the estimated STFT representations of the nearend speech respectively, $|*|$ is the modular operation, ω is a controlling scalar with value 0.3, P_{vad} is the estimated VAD state, P is the groundtruth of the VAD of nearend speech generated by the WebRTC-VAD, and β is a controlling scalar with value 0.06.

For personalized AEC, the $loss_{AEC}$ item can be reformulated as

$$loss_{pAEC} = \omega\mathcal{L}(S,\hat{S}_1) + (1-\omega)\mathcal{L}(S_t,\hat{S}_2) \quad (4)$$

where S_t denotes the groundtruth STFT representations of the nearend target speech.

2.5. Data cleaning

In this paper, the real farend single-talk recordings are used as the echo data to simulate the training data. However, there are a small amount of noisy data, which have the presence of nearend speakers and would lead to nearend speech distortion. To alleviate this problem, a simple but effective data cleaning strategy is employed. We first simulate about 2000 hours of training data by mixing the echo and clean speech, and train a general AEC model. The real farend single-talk recordings are then fed into the model and the recordings with high residual energy will be identified as noisy data. This cleaning strategy is performed iteratively until the amount of noisy data is below a threshold. The clean speech and echo data are selected from the DNS Challenge[2] and LibriSpeech dataset[3], and the AEC Challenge training farend single-talk clips respectively. The RIR set from DNS Challenge is applied to simulate reverberation near-end signal. Since the echo data already contains near-end noise, no noise is applied during data generation.

3. EXPERIMENTS AND RESULTS

We use the STFT block with window length 20 ms and hop length 10 ms. All network layers are configured as causal and the BatchNorm is only performed during training stage. As shown in Table 1, the proposed general AEC model obtains a 0.420 gain on subjective-MOS and a 0.174 gain on WAcc ratio, compared with baseline. The proposed model can further promote the subjective-MOS by 0.011 when SAM is introduced for personalized AEC. The proposed model ranks first in the ICASSP 2023 AEC Challenge Track 1 and Track 2, which proves the robust performance of the proposed framework. The model size of the general and personalized AEC models are 1.48 M and 1.64 M respectively. Tested on a CPU of 2.30GHz, the real time factor of the general and personalized AEC models with Pytorch implementation are about 0.283 and 0.310 respectively.

4. CONCLUSION

This paper introduced our hybrid LAEC and two-stage neural AEC model in the ICASSP 2023 AEC Challenge. The proposed method provides state-of-the-art performance in the Track 1 and Track 2.

5. REFERENCES

[1] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, E. Indenbom, N. Ristea, J. Guzhvin, H. Gamper, S. Braun, and R. Aichner, "Icassp 2023 acoustic echo cancellation challenge," in *ICASSP*, 2023.

[2] J. Yu, Y. Luo, H. Chen, R. Gu, and C. Weng, "High fidelity speech enhancement with band-split rnn," *arXiv preprint*, 2022.

[3] X. Le, Z. Hou, L. Chen, C. He, Y. Guo, C. Chen, X. Xia, and J. Lu, "Personalized speech enhancement combining band-split rnn and speaker attentive module," in *ICASSP*, 2023.

[4] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint*, 2020.

[5] G. Zhang, L. Yu, C. Wang, and J. Wei, "Multi-scale temporal frequency convolutional network with axial attention for speech enhancement," in *ICASSP*. IEEE, 2022.

[6] 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*. IEEE, 2022.

[2]https://github.com/microsoft/DNS-Challenge

[3]http://www.openslr.org/12/