ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747347

EXPLORE RELATIVE AND CONTEXT INFORMATION WITH TRANSFORMER FOR JOINT ACOUSTIC ECHO CANCELLATION AND SPEECH ENHANCEMENT

Xingwei Sun, Chenbin Cao, Qinglong Li, Linzhang Wang, Fei Xiang

Xiaomi Corporation, Beijing, China

ABSTRACT

This paper proposes a joint acoustic echo cancellation (AEC) and speech enhancement method with adaptive filter and deep neural network (DNN) model. A partitioned block adaptive filter is adopted for linear AEC followed by a convolutional neural network and transformer based model to suppress the residual echo, noise, and reverberation. The DNN model has three modules: encoder, dual-path transformer (DPT) and decoder. The encoder is adopted to explore the potential relationships of far-end and near-end signals with the attention mechanism of transformer. The DPT module is further used to explore context information in both time and frequency dimension. The attention mask is used in transformer to realize real-time process. The complex spectra mask is finally estimated by the decoder to recover the target speech. Our proposed DNN model is trained on the ICASSP 2022 AEC Challenge datasets and placed fourth in the challenge with satisfactory performance on subjective and word acceptance rate evaluation.

Index Terms— acoustic echo cancellation, speech enhancement, deep neural network, transformer

1. INTRODUCTION

The adaptive filter [1] is generally used for acoustic echo cancellation (AEC) while it cannot remove the annoying nonlinear components. Many methods have been proposed to address nonlinear echo cancellation [2, 3, 4] with expensive storage and computation cost as well as slow convergence. Traditional methods employ spectral enhancement techniques to residual echo suppression (RES) [5, 6]. However, due to the high non-stationarity of echo, the underling under- and overestimation leads to lots of residual echo, musical noise, and speech distortion.

In recent years, deep neural network (DNN) based methods have shown great advantages in speech enhancement [7]. Recurrent neural network and convolutional recurrent neural network based architectures are proposed for joint residual echo and noise suppression [8, 9]. The complex neural networks have shown significant benefits for complex spectra modeling in speech enhancement [10] and AEC tasks[11]. A gate complex convolution recurrent neural network is also proposed to suppress the residual echo, late reverberation and environmental noise simultaneously [12].

Inspired by the outstanding performance of transformer structure applied in speech separation [13] and speech recognition [14] tasks, we proposed an encoder-decoder framework consisted of convolutional neural network (CNN) and transformer layers to estimate complex mask for joint residual echo, noise and late reverberation suppression. To avoid confusions, the transformer in this paper refers specially to the encoder part of transformer as proposed in [15]. In our proposed DNN model, the transformer layers are adopted to explore the related features of far-end reference signal and near-end mix signal in different potential feature domain due to the attention mechanism in multi-head attention with reference feature as the query and mix feature as key and value. The CNN layers are adopted to map the features into different potential feature domain. Further, the dual-path transformer (DPT) module is adopted to explore the context information in both time and frequency dimension. Finally, the complex mask is estimated to recover the target speech. By training the model with the datasets from the ICASSP 2022 AEC Challenge [16], our system achieved outstanding performance compared with the baseline method and placed fourth in the challenge.

2. AEC SIGNAL MODEL

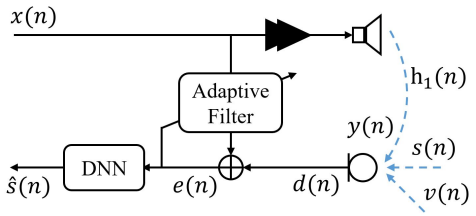

Fig. 1. Diagram of joint AEC and speech enhancement.

The signal model of echo cancellation and speech enhancement system is illustrated in Fig.1. The far-end reference signal $x(n)$ is amplified by a power amplifier and played by a loud speaker, where nonlinear components appear. The echo replica $y(n) = x(n - \triangle) \star h_1(n)$ is picked up by a

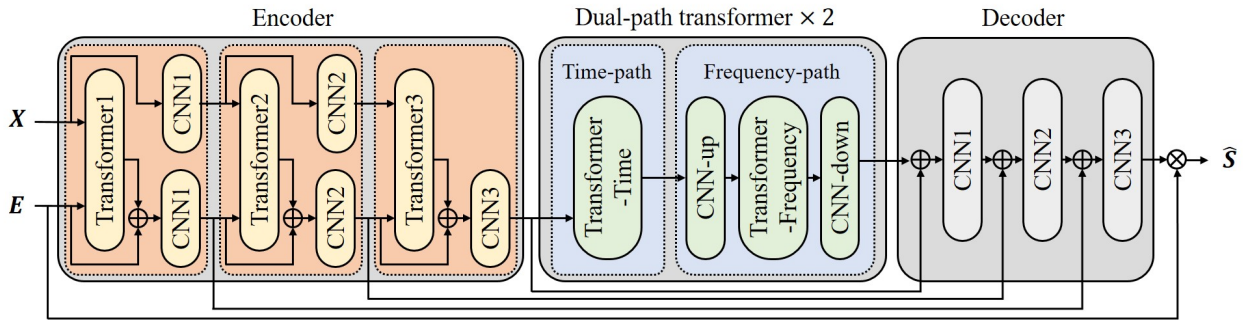

Fig. 2. Framework of the DNN model for joint RES and speech enhancement.

microphone along with the observed near-end speech $s(n)$ and noise signal $v(n)$. The microphone signal received signal $d(n)$ can be described as:

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

Here, $\star$ presents linear convolution, n denotes the discrete time index and $\triangle$ denotes the time delay between far-end signal and echo replica, $h_1(n)$ is the room impulse response (RIR) between microphone and loud speaker. The output signal $e(n)$ after linear AEC process can be written as:

$$e(n) = d(n) - x(n) \star \hat{h}_1(n). \tag{2}$$

Here, $\hat{h}_1(n)$ is the estimate of $h_1(n)$ with adaptive filter. $e(n)$ is still referred as near-end mix signal afterwards. Further considering reverberation, the observed speech signal $s(n)$ at the microphone can be divided into two part: direct sound with early reflection $s_d(n)$ and the late reverberation $s_l(n)$. In our proposed DNN model training process, $s_d(n)$ is generated with the first 50 ms of $h_1(n)$ and used as target speech.

3. DNN MODEL

After the cancellation of linear echo signal with adaptive filter, we proposed a DNN model to suppress the residual echo, noise and late reverberation in short time Fourier transformation (STFT) domain with a complex mask estimation. As depicted in Fig.2, our system consists of three stages: encoder, DPT and decoder. First, the complex spectra of far-end reference signal and near-end mix signal are processed by the encoder to explore the related features of these two signals in different potential feature domain. Then, the encoded features are fed into DPT module to explore the context information in both time and frequency dimension. Finally, the decoder is used to estimate the complex mask for recovering complex spectra of target speech.

3.1. CNN and transformer based encoder

The encoder module consists of three blocks: the first two have one transformer layer and two CNN layers, and the last one has no CNN layer in the reference signal branch, as shown in encoder part of Fig.2. In each block, the transformer layer is used to explore the relative features of the reference and mix signals. In our proposed model, we adopt an improved transformer layer in which there are mainly two modification compared with the original transformer structure. First, the masked multi-head attention [14] is used to control the usage length of context information, for example, formulating a causal transformer layer without future information. The other one is using LSTM layer instead of fully connected layer [13] in the feed forward network to model the sequence order information without positional encodings. The improved transformer can be formulated as follows:

$$Q_i = Z^Q W_i^Q, K_i = Z^K W_i^K, V_i = Z^V W_i^V, \tag{3}$$

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax}(\frac{\text{mask}(Q_i K_i^T)}{\sqrt{d}})V_i, \tag{4}$$

$$\text{MultiHead} = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \tag{5}$$

$$\text{Mid} = \text{LayerNorm}(Z^V + \text{MultiHead}), \tag{6}$$

$$\text{FFN} = \text{ReLU}(\text{LSTM}(\text{Mid}))W + b, \tag{7}$$

$$\text{Output} = \text{LayerNorm}(\text{Mid} + \text{FFN}). \tag{8}$$

Here, $i \in [1, h]$ is the head index. $Z^Q, Z^K, Z^V \in R^{m \times d}$ are the input features with length m and dimension d, $Q_i, K_i, V_i \in R^{m \times d/h}$ are the mapped queries, keys and values. $W^Q, W^K, W^V \in R^{d \times d/h}$ and $W^O \in R^{d \times d}$ are parameter matrices. The $mask$ operator in multi-head attention denotes the attention mask to control the length of context information to be used. FFN denotes the output of the feed forward network. In our proposed model, the spectrogram feature of the far-end reference signal is used as Z^Q, and that of the near-end mix signal is used as Z^K and Z^V.

The CNN layers are used to map the features of the two signals into different potential feature domain. Note that the input features in mix signal branch are added with the relative features before feeding into the CNN layer. The CNN layers of the reference and mix signal branches in the same

block have the same structure, with each CNN layer consists of one convolution layer followed by batch normalization and parametric rectified linear unit (PReLU) activation.

3.2. Dual-path transformer network

After the encoder module, the encoded features are sequentially fed into two DPT modules each with a time-path and frequency-path process, as shown in the DPT part of Fig.2. The time-path process includes a time-domain transformer layer to explore the context information in time dimension. The transformer structure in the time-path process is the same with the improved transform in encoder module, in which the time dimension is used as the sequence dimension and the channel and frequency dimension as feature dimension. The reshape operators are applied before and after transformer layer between the features of $Z \in R^{B\times C\times T\times F}$ and $Z \in R^{B\times T\times FC}$, where B, C, T, F are the dimension of batch size, channel, time and frequency, $FC = F * C$ is the combined dimension of frequency and channel.

The frequency-path process includes a frequency-domain transformer layer and two CNN layers. The transformer layer is used to explore the context information in frequency dimension. Thus, the frequency dimension is used as the sequence dimension with the time dimension combined into batch size and the channel dimension as feature dimension. The reshape operators between the features of $Z \in R^{B\times C\times T\times F}$ and $R^{BT\times F\times C}$ are applied, where $BT = B * T$ is the combined dimension of batch size and time. There is no mask operator in the multi-head attention of this transformer. However, the bidirectional LSTM layer is adopted in the feed forward network. Different from the time-path process, the frequency-path process also has two CNN layers to expand or reduce the feature dimension (the original channel dimension). In each of these two CNN layers, only one convolution layer is used.

3.3. Decoder

The decoder includes three CNN layers and skip connections with encoder, as shown in the decoder part of Fig.2. The CNN layers map the features in potential feature domain back to the STFT domain and output the complex mask. One transposed convolution layer followed by batch normalization and PReLU activation is used in each CNN layer. The features in the same potential feature domain of encoder and decoder are added before feeding into the next CNN layer with the skip connection. Finally, the complex mask is apply to the near-end mix signal to recover the estimation of target speech.

4. EXPERIMENT SETUP

4.1. Data preparation

The datasets provided by the organizer consist of real and synthetic recordings whose delay between far-end signal and echo replica are long or varying [16]. The partitioned block frequency-domain adaptive filter [17] method is adopted for real-time delay estimation. The far-end signal is shifted to approximately align with near-end signal. The partitioned block Kalman adaptive filter is adopted for linear echo cancellation [18, 19]. In order to re-converge fast and keep the Kalman filter from divergence when echo path changes, a two-path mechanism [20] is introduced, where a secondary variable-step-size based normalized least mean square adaptive filter works in-parallel with the primary Kalman filter.

All data used for model training is filtered and aligned using the above algorithm. For the synthetic data, the far-end reference, near-end microphone and speech signals can be used directly. For the real data, we use only far-end single talk data to generate training data by randomly select the far-end signal as reference signal and the microphone received signal as echo signal and randomly add near-end speech and noise signals with and without reverberation. The speech, noise and RIR signals from ICASSP 2022 deep noise suppression challenge [21] are used.

4.2. Input/output features and loss function

The complex spectra after STFT of far-end reference signal and that of the near-end mix signal after linear echo cancellation are denoted as $X(k,l)$ and $E(k,l)$. We use the compressed complex spectra [12] as the input features and training target. The compress and uncompress process can be described as:

$$A^c(k,l) = |A(k,l)|^{0.5} e^{j\angle A(k,l)}, \tag{9}$$

$$A^u(k,l) = |A(k,l)|^2 e^{j\angle A(k,l)}. \tag{10}$$

Here, $A(k,l)$ is the complex spectra to be processed with $|A(k,l)|$ and $\angle A(k,l)$ according to its magnitude and phase, k and l are the frequency bin and time frame index. The real and imaginary part of the compressed complex spectra $X^c(k,l)$ and $E^c(k,l)$ are used as two channel when feeding into DNN model. The estimated compressed complex spectra of the target speech is obtained by applying the mask (output of DNN model) to $E^c(k,l)$, which can be written as:

$$\hat{S}^c(k,l) = |E^c(k,l)|\tanh(|M(k,l)|)e^{\angle E^c(k,l)+\angle M(k,l)}. \tag{11}$$

Here, the operator $tanh$ is used to restrict the mask magnitude in $[0, 1]$. Finally, the estimated uncompressed complex spectra $\hat{S}^u(k,l)$ is used to recover the time domain speech after an inverse STFT process.

The loss function we used to train the DNN model is mean square error (MSE) of magnitude spectra and complex spectra, which can be written as (with (k,l) omitted):

$$\begin{aligned} J =& \text{MSE}(|\hat{S}^c|, |S_d^c|) + \text{MSE}(\mathcal{R}(\hat{S}^c), \mathcal{R}(S_d^c)) \\ &+ \text{MSE}(\mathcal{I}(\hat{S}^c), \mathcal{I}(S_d^c)). \end{aligned} \tag{12}$$

Here, S_d^c is the compressed complex spectra of the target speech (s_d in Section 2). $\mathcal{R}$ and $\mathcal{I}$ are the operators to obtain the real and imaginary part of complex number.

4.3. Model configuration and training setup

The hyper parameters of each layer are shown in Table 1. The hyper parameters of the CNN layer are presented as *channel number, kernel size and stride size of (transposed) convolution layer in frequency-time dimension*. No future frame is used in CNN layers with proper padding process in time dimension. The structures of transformer layers have been detailed in previous sections. In the masked multi-head attention module, we used 49 previous frames and one future frame by generating applicable attention mask matrix. The hyper parameters of transformer layer are presented as *input feature dimension, head number and hidden units number of LSTM layer*.

module	layer	hyper parameter
encoder	transformer1	960, 4, 256
	CNN1	8,(5, 2),(2, 1)
	transformer2	1920, 4, 256
	CNN2	16,(7, 2),(3, 1)
	transformer3	1280, 8, 256
	CNN3	32,(5, 2),(2, 1)
DPT	transformer-time	1280, 4, 256
	CNN-up	256, (1,1),(1,1)
	transformer-frequency	256, 2, 256
	CNN-down	32, (1,1),(1,1)
decoder	CNN1	16,(5, 2),(2, 1)
	CNN2	8, (7, 2),(3, 1)
	CNN3	2, (5, 2),(2, 1)

Table 1. The hyper parameters of each layer.

During training and testing, all time-domain signals are framed by a Hanning window with 20ms window length and 10ms hop size. As the sample rate is 48k Hz, a 960-point STFT is applied to each time frame to produce the complex spectra. The frequency dimension is 480 without the direct current component. Our model is trained 350 epochs using the Adam optimizer with learning rate of 1e-4. We used 8 utterances as a batch and each epoch has 2000 iterations.

5. EVALUATION RESULTS

5.1. Simulated test dataset

To evaluate the performance of our proposed method, we generated a simulated test dataset with the far-end single talk signals in the test dataset of ICASSP 2022 AEC challenge. We generated 500 utterances with each 10 seconds including three different scenarios in the same way with the usage of real data in model training as described in section 4.1. The perceptual evaluation of speech quality (PESQ) evaluation results of double talk and near-end single talk scenarios and echo return loss enhancement (ERLE) of far-end single talk scenario are shown in Table 2. The results shows that our proposed method obtained outstanding performance and obviously better than the baseline method.

Method	DT PESQ	ST NE PESQ	ST FE ERLE
orignal	1.09	1.09	-
baseline	1.77	2.60	-39.45
ours	1.93	3.14	-46.68

Table 2. Evaluation results of simulated test dataset.

5.2. Blind test dataset

Table 3 shows the evaluation results of the blind test dataset released from the organizer. The results show that our method significantly outperforms the baseline in both subjective and word acceptance rate (WAcc) evaluations. Finally, our proposed system placed fourth in this challenge. However, the performance of our proposed method still have gap compared with the team placed first.

Team	ST NE MOS	ST FE MOS	DT ECHO DMOS	DT other DMOS	WAcc	Final
1st	4.403	4.811	4.754	4.431	0.814	0.883
ours	4.226	4.704	4.667	4.067	0.772	0.838
baseline	4.152	4.563	4.122	3.563	0.659	0.752

Table 3. Evaluation results of blind test dataset.

5.3. Computation complexity

As one future frame is used in transformer layers, the total algorithmic latency of our system is 40ms. The real time factor (RTF) is 0.3315, tested on a laptop computer with Intel(R) Core(TM) i5-11300H@3.10GHz. The total number of parameters of our DNN model is 51.48 millions.

6. CONCLUSION

We proposed a real-time joint AEC and speech enhancement system with adaptive filter and DNN model. In the DNN model, the transformer in the encoder is used to explore the relative information between the far-end reference signal and near-end mix signal with these two signal as input while the DPT model explores the context information in both time and frequency dimension. Finally, the complex spectra mask is estimated by decoder to recover target speech. Our proposed method obtained satisfactory performance with PESQ and ERLE evaluation on a simulated dataset and subjective and WAcc evaluation on a real dataset.

7. REFERENCES

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

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

[3] Fabian Kuech and Walter Kellermann, "Partitioned block frequency-domain adaptive second-order volterra filter," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 564–575, 2005.

[4] Fabian Kuech and Walter Kellermann, "Nonlinear residual echo suppression using a power filter model of the acoustic echo path," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. IEEE, 2007, vol. 1, pp. I–73.

[5] Seon Joon Park, Chom Gun Cho, Chungyong Lee, and Dae Hee Youn, "Integrated echo and noise canceler for hands-free applications," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 49, no. 3, pp. 188–195, 2002.

[6] Yun-Sik Park and Joon-Hyuk Chang, "Frequency domain acoustic echo suppression based on soft decision," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 53–56, 2008.

[7] DeLiang Wang and Jitong 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.

[8] Jean-Marc Valin, Srikanth Tenneti, Karim Helwani, Umut Isik, and Arvindh Krishnaswamy, "Low-complexity, real-time joint neural echo control and speech enhancement based on percepnet," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7133–7137.

[9] 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.

[10] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Interspeech*, 2020, pp. 2472–2476.

[11] 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 *Interspeech*, 2021, pp. 4758–4762.

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

[13] Jingjing Chen, Qirong Mao, and Dong Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[14] Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7829–7833.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] Ross Cutler, Ando Saabas, Tanel Parnamaa, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sorensen, and Robert Aichner, "ICASSP 2022 acoustic echo cancellation challenge," in *ICASSP 2022*.

[17] J-S Soo and Khee K Pang, "Multidelay block frequency domain adaptive filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 373–376, 1990.

[18] Gerald Enzner and Peter Vary, "Frequency-domain adaptive kalman filter for acoustic echo control in hands-free telephones," *Signal Processing*, vol. 86, no. 6, pp. 1140–1156, 2006.

[19] Fabian Kuech, Edwin Mabande, and Gerald Enzner, "State-space architecture of the partitioned-block-based acoustic echo controller," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1295–1299.

[20] Feiran Yang, Gerald Enzner, and Jun Yang, "Frequency-domain adaptive kalman filter with fast recovery of abrupt echo-path changes," *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1778–1782, 2017.

[21] Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sergiy Matusevych, Sebastian Braun, Emre Sefik Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, and Robert Aichner, "ICASSP 2022 deep noise suppression challenge," in *ICASSP*, 2022.