INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Acoustic Echo Cancellation using Deep Complex Neural Network with Nonlinear Magnitude Compression and Phase Information

Renhua Peng[1,2], *Linjuan Cheng*[1,2], *Chengshi Zheng*[1,2], *Xiaodong Li*[1,2*]

[1]Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
{pengrenhua, chenglinjuan, cszheng, lxd}@mail.ioa.ac.cn

Abstract

This paper describes a two-stage acoustic echo cancellation (AEC) and suppression framework for the INTERSPEECH2021 AEC Challenge. In the first stage, four parallel partitioned block frequency domain adaptive filters are used to cancel the linear echo components, where the far-end signal is delayed 0ms, 320ms, 640ms and 960ms for these four adaptive filters, respectively, thus a maximum 1280 ms time delay can be well handled in the blind test dataset. The error signal with minimum energy and its corresponding reference signal are chosen as the input for the second stage, where a gate complex convolutional recurrent neural network (GCCRN) is trained to further suppress the residual echo, late reverberation and environmental noise simultaneously. To improve the performance of GCCRN, we compress both the magnitude of the error signal and that of the far-end reference signal, and then the two compressed magnitudes are combined with the phase of the error signal to regenerate the complex spectra as the input features of GCCRN. Numerous experimental results show that the proposed framework is robust to the blind test dataset, and achieves a promising result with the P.808 evaluation.

Index Terms: acoustic echo cancellation, INTERSPEECH AEC Challenge, deep learning

1. Introduction

Acoustic echo cancellation (AEC) is very important in hands-free teleconference applications. Traditional linear AEC algorithms [1] can achieve promising performance when the acoustic echo path is linear and time invariant. However, the performance of these existing linear AEC algorithms may be degraded greatly in many realistic applications due to many factors, such as a large dynamic time delay of the reference signal, the nonlinear behaviors of the loudspeaker, time variations of the acoustic echo path, double-talk scenarios and so on. All the above mentioned problems need to be handled for practical hand-free teleconference systems. Though many efforts have been made to increase the convergence rate of adaptive filtering algorithms and improve their robustness against double-talk scenarios [2], [3], [4], there still exists a large gap in realistic applications.

In recent years, it becomes popular to combine the linear AEC algorithm with nonlinear post processing (NLP) using deep neural networks (DNN) in this area [5], [6], [7], [8]. In [8], we use the partitioned block frequency-domain least mean square (PBFDLMS) algorithm for the ICASSP 2021 AEC Challenge [9], which can achieve a good balance between the computational complexity and the algorithmic delay. For this AEC Challenge, we also adopt the PBFDLMS algorithm to cancel the linear echo components first, and propose to use a convolutional recurrent neural network for the non-linear post processing. In [7], the original far-end reference signal and the near-end microphone signal are chosen to train the NLP network, which can improve the robustness when the time delay estimation fails. However, we find that using the time aligned far-end reference signal and the near-end microphone signal to train the NLP network can further improve the performance of NLP network. Thus, in this paper, we use the time-aligned far-end reference signal to train the NLP network.

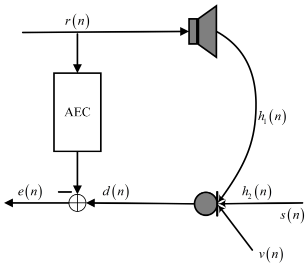

Figure 1: *Diagram of a linear AEC system.*

The remainder of this paper is organized as follows: Section 2 introduces the signal model and formulates the problem. In Section 3, the overall framework of the proposed GCCRN model is described in detail and the experimental setup is presented in Section 4. In Section 5, some subjective and objective experimental results are presented to show the effectiveness of the proposed framework. Section 6 presents some conclusions.

2. AEC model

Fig. 1 plots the diagram of a linear AEC system. The far-end reference signal $r(n)$ is played by the loudspeaker and then recorded by a microphone via the acoustic echo path $h_1(n)$. The near-end speech signal $s(n)$ is also recorded by the microphone via the acoustic path $h_2(n)$. Accordingly, the microphone signal can be modelled as:

$$d(n) = r(n)*h_1(n) + s(n)*h_2(n) + v(n), \tag{1}$$

where $*$ denotes convolution, n represents the discrete-time index, $s(n)$ is the near-end speech signal. $h_1(n)$ and $h_2(n)$ are the room impulse responses (RIR) between the microphone and the loudspeaker, and the near-end speech, respectively. $v(n)$ is the additive environmental noise. In [8], only the clean near-end speech without reverberation is considered, which may not be true in realistic scenarios. In this paper, we consider the acoustic echo, the near-end speech reverberation, and the additive environmental noise together, and aim to recover the early reflected

[*] corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-2022

near-end speech signal $s_e(n)$, which can be given by:

$$s_e(n) = \sum_{i=0}^{L_e} s(n-i)\, h_2(i) \tag{2}$$

where L_e ranges from $0.05f_s$ to $0.08f_s$ with f_s the sampling rate. The main reason that we preserve the first 50 ms to 80ms early reflected speech signal is that these early reflected sounds do not degrade speech quality significantly.

It's well known that linear AEC algorithms can cancel the echo signal $r(n)*h_1(n)$ from $d(n)$ with the a priori information of the reference signal $r(n)$ under linear and time-invariant system assumption. In this paper, we use the same adaptive AEC algorithm as in [8], called partitioned block frequency domain least mean square (PBFDLMS) algorithm, to cancel the linear echo components.

3. DNN Model

Though PBFDLMS can cancel the echo signal to a certain extent, some residual echo components are inevitable due to under-modeling of $h_1(n)$, nonlinear distortion, and so on. We propose to use a gate complex convolutional recurrent network (GCCRN) to suppress residual echo, late reverberation and additive environmental noise together in the following.

3.1. Network structure

As shown in Fig. 2, the network structure has 5 gate convolutional encoder layers with kernels of size $(2, 3)$ in time and frequency dimensions, and 5 symmetric gate de-convolutional layers for both the real and imaginary parts, where 'GConv2d' represents the gate convolutional encoder layers and 'GDeconv2d' represents the gate de-convolutional layers, respectively. The convolution kernels move with a stride of $(1, 2)$, i.e. downsample the features along the frequency axis efficiently. The number of output channels for each layer in encoder part is $(16, 32, 64, 128, 256)$. The skip connections can be implemented by simply adding the encoder outputs to the decoder inputs, which can reduce the number of decoder input channel compared to concatenating. By doing so, the number of input channels for each layer in decoder part is $(256, 128, 64, 32, 16)$. In [10], it has been shown that adding a trainable channel-wise scaling and bias in the add-skip connections can improve the performance at negligible additional cost. We adopt these improvements and add a convolution layer with $(1, 1)$ kernel size and $(1, 1)$ stride in the add-skip connections. Two layers of long short term memory (LSTM) are stacked between the encoder and decoder to explore the temporal relations of speech signals [11]. For each LSTM layer, we group the wide fully connected LSTM into two disconnected parallel LSTMs, still yielding the same forward information flow as shown in Fig. 2.

3.2. Input/Output features

The short time Fourier transformation (STFT) of far-end reference signal and that of the error signal are denoted as $R(k, l)$ and $E(k, l)$, respectively. We propose to use the compressed complex spectra as the input features to train the proposed GCCRN model, where we have found that compressed complex spectral features are quite important in speech de-reverberation [12], which is given by:

$$\widehat{R}_r(k,l) = |R(k,l)|^{1/2} \cos(\angle E(k,l)) \tag{3}$$

$$\widehat{R}_i(k,l) = |R(k,l)|^{1/2} \sin(\angle E(k,l)) \tag{4}$$

$$\widehat{E}_r(k,l) = |E(k,l)|^{1/2} \cos(\angle E(k,l)) \tag{5}$$

$$\widehat{E}_i(k,l) = |E(k,l)|^{1/2} \sin(\angle E(k,l)) \tag{6}$$

where $|R(k,l)|$ and $|E(k,l)|$ are the magnitude of $R(k,l)$ and $E(k,l)$, respectively. $\angle E(k,l)$ is the phase of the error signal. Both the real and imaginary parts of the compressed spectra are concatenated to form a tensor of dimension $(4, T, 161)$, which means that the input channel C_{in} of the GCCRN model in Fig. 2 is 4. T represents the frame number of each training samples. Note that we construct the real and imaginary parts of the far-end reference signal using the phase information of the error signal, which improves robustness of the NLP to the time difference between the far-end reference signal and the microphone signal.

The output features of the proposed GCCRN model are the real and imaginary parts of the compressed complex spectrum of $s_e(n)$, which is given by:

$$\widehat{O}_r(k,l) = |S_e(k,l)|^{1/2} \cos(\angle S_e(k,l)) \tag{7}$$

$$\widehat{O}_i(k,l) = |S_e(k,l)|^{1/2} \sin(\angle S_e(k,l)) \tag{8}$$

where $S_e(k,l)$ is the STFT of $s_e(n)$.

Both $\widehat{O}_r(k,l)$ and $\widehat{O}_i(k,l)$ are formed a tensor of dimension $(1, T, 161)$, which means that the output channel C_{out} of the GCCRN model in Fig. 2 is 1. Note that when constructing the time domain signal of the enhanced speech, the real and imaginary outputs of the GCCRN model should be uncompressed beforehand and then combine with the phase information.

4. Experimental setup

4.1. Datasets and preparation

Two datasets are provided through the challenge [13], where one is synthesized data and the other is real-recording data. Because there is no clean speech in real recording dataset, it is not suitable for constructing the training datasets without additional processing. However, we find that the clean speech in synthesized dataset cannot be used directly either. According to the system description of the synthesize data generation [13], the amplitude of the near-end clean speech may be different from that in the near-end microphone signal due to having the automatic gain control (AGC). So gain normalization should be made on the clean speech beforehand, and the normalized gain can be obtained by:

$$\alpha = \sum_{n=0}^{N-1} (d(n)\, s(n)) / \sum_{n=0}^{N-1} s^2(n) \tag{9}$$

where $d(n)$ and $s(n)$ are the near-end microphone signal and near-end clean speech signal in synthesized dataset, respectively, and N is the number of samples in each audio clips. Then $\alpha s(n)$ is used as the target clean speech.

To improve the generalization capability of GCCRN, the training dataset includes many languages, such as Korean [14], Mandarin [15], English [16], Spanish [17], Japanese [18] and Germany [19]. All the audio clips are re-sampled to 16 kHz and then separated into two parts, where one is for the far-end reference signal, and the other for the near-end clean speech source signal. Both the echo paths $h_1(n)$ and the reverberation paths $h_2(n)$ are gathered from [20], and the noise corpus is provided by [21]. When training the model, the signal-to-noise

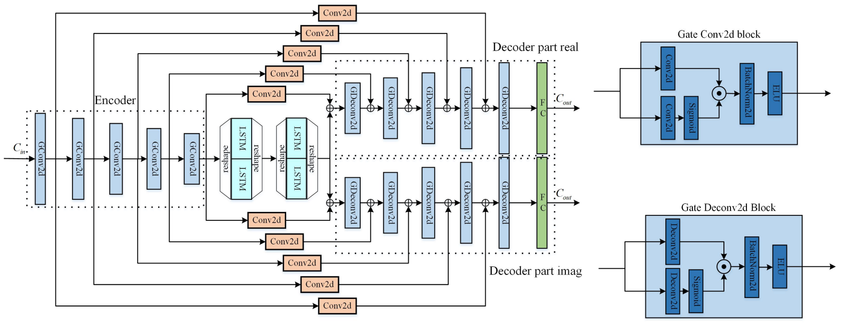

Figure 2: *The architecture of proposed GCCRN model.*

ratio (SNR) ranges from -5 dB to 10 dB and the source-to-echo ratio (SER) ranges from -15 dB to 15 dB.

It should be noted that all the far-end reference signal $r(n)$, the near-end clean speech signal $s(n)$, the echo path $h_1(n)$, the reverberation path $h_2(n)$ and additive environmental noise $v(n)$ are randomly selected from its corresponding dataset independently. After creating the near-end microphone signal, the far-end reference signal and the clean speech signal datasets, all the audio pairs of far-end reference signal and the near-end microphone speech are pre-processed by the PBFDLMS algorithm to obtain the error signal, where the number of sub-blocks used in the PBFDLMS algorithm is 32 with the block length 160. At last, a total of 40000 segments of training speech pairs and 4000 segments of testing and validation speech pairs are prepared for network training and validation.

4.2. Training setup

All the time-domain signals of the far-end reference speech, the near-end microphone speech and the clean target speech are transformed into the frequency domain by applying a Hanning window to a frame of 320 consecutive samples with the frame shift of 160 samples. In the proposed framework, we do not use any future information to further improve the performance, and thus the algorithmic delay is 30 ms according to the rule of this challenge when the sampling rate f_s is 16 kHz.

The network is trained to optimize the minimum square error (MSE) loss function between the network output and the near-end clean speech with the Adam optimizer. The number of trainable parameters of the proposed GCCRN model is $10204524 \approx 10.2$M. 200 epochs are set in the training stage and the best model parameters having minimum validation loss is saved.

5. Results

5.1. Reference delay

In the pre-processing stage of PBFDLMS and the training stage of the GCCRN model, we assume that the far-end reference signal and the near-end microphone signal are time aligned. We first study the impact of the time difference between the far-end reference signal and the microphone signal on the AEC algorithm and the GCCRN model performance using PESQ scores [22], Fig. 3 presents the results under different time delay of the far-end reference signal. The positive value of delay time in Fig. 3 represents that the far-end reference signal is ahead, and the negative value represents that the far-end reference signal is postponed.

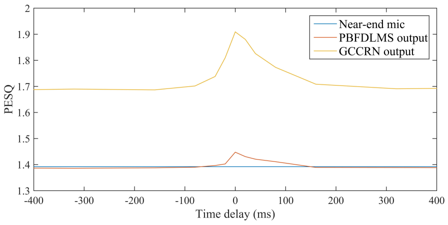

Figure 3: *PESQ performance versus reference time delay.*

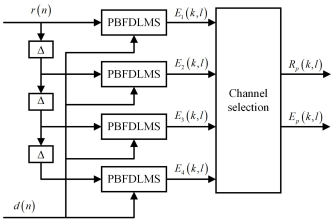

Figure 4: *Channel selection of error signal.*

It can be seen from Fig. 3 that the PBFDLMS algorithm is sensitive to the time difference, which improves the PESQ score when the time delay changes from -20 ms to 150 ms. Both the PBFDLMS algorithm and the proposed GCCRN model can achieve the best PESQ performance when the far-end reference signal and the near end microphone are time aligned, and the performance is greatly reduced when increasing the time difference.

5.2. Blind dataset processing

For the blind test dataset released from the organizer, the time delay between the far-end reference signal and near-end microphone has a large dynamic range, and thus it is important to compensate the time difference beforehand. To alleviate the computational complexity of the time delay estimation function used in [8], and to reduce the time delay estimation error, we propose to use four parallel PBFDLMS blocks with pre-delayed far-end reference signal, as shown in Fig. 4, where the time delay estimation function is also included in the PBFDLMS block.

The pre-delayed value Δ is set to 320 ms, and the searching time window in each time delay estimation function is also set to 320 ms, which means a total of 1280 ms time delay can be covered and well handled in the blind test dataset.

It is well-known that when the far-end reference signal and

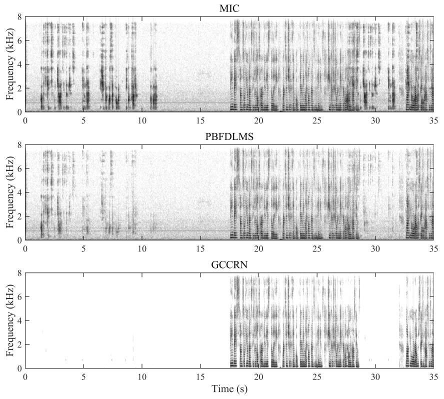

Figure 5: *A double-talk utterance in blind test dataset. (up): original microphone, (middle): enhanced speech processed by PBFDLMS, (bottom): enhanced speech processed by GCCRN.*

Table 1: *MOS results in INTERSPEECH2021 Challenge*

	Baseline	Proposed
ST NE MOS	4.18	**4.26**
ST FE Echo DMOS	3.82	**4.34**
DT Echo DMOS	4.04	**4.36**
DT Other DMOS	3.45	**4.23**

near-end microphone is time aligned, the echo signal in near-end microphone will be attenuated most, which means we can select the best error signal and far-end reference signal for GCCRN model processing according to the energy of the error signal. The best channel p is selected as:

$$p = \arg\min_{i} |E_i(k,l)|^2 \tag{10}$$

and $R_p(k,l)$ denotes the STFT of $r(n - p\Delta)$.

To validate the effect of the two stages intuitively, we plot the spectrograms of the near-end microphone signal, which is taken from the blind test dataset, the selected error signal processed by the four PBFDLMS blocks and the output of the proposed GCCRN model. One can see from Fig. 5 that the acoustic echo can be partially suppressed by the PBFDLMS algorithm, while the near-end speech remains undistorted, and the GCCRN model can well suppress the residual echo. Besides the residual echo, the proposed GCCRN model is also able to suppress the environmental noise and late reverberation because the training target is the early-reflected near-end speech..

5.3. Objective measurement results

Table. 1 summarizes the mean opinion scores (MOS) conducted by the organizers according to this Challenge rules [23]. One can see that the proposed framework gets a large improvement compared to the baseline system. Tables. 2 and 3 present the degradation MOS and the echo MOS results, respectively, using the DECMOS tool provide by the organizers [1], where 'Original' represents the original microphone signal, 'PBFDLMS' represents the error signal obtained by the PBFDLMS algorithm, and 'DCGRU' represents the algorithm proposed in [8], 'MOV' represents the movement scenario. One can see that the proposed algorithm in this paper achieves the lowest degradation MOS score and the highest echo MOS score in different echo scenarios. In the near-end single-talk scenario, the proposed framework shows negligible degradation in echo MOS score.

Table 2: *Degradation MOS results*

	Original	PBFDLMS	DCGRU	Proposed
ST NE	3.73	3.74	3.76	**3.71**
ST FE	5.00	5.00	5.00	5.00
ST FE MOV	5.00	5.00	5.00	5.00
DT	3.58	3.48	3.30	**3.00**
DT MOV	3.56	3.48	3.24	**2.91**

Table 3: *Echo MOS results*

	Original	PBFDLMS	DCGRU	Proposed
ST NE	**4.98**	4.96	4.95	4.94
ST FE	2.06	2.34	3.19	**4.23**
ST FE MOV	1.98	2.10	2.82	**4.11**
DT	2.46	2.73	3.34	**3.90**
DT MOV	2.54	2.70	3.28	**3.83**

5.4. Computation complexity

We run our framework code on a laptop computer with Intel(R) Core(TM) i5-4300 CPU @1.9GHz, the computational time per frame of the PBFDLMS algorithm is 0.08ms, and the four parallel structures take a total of 0.32 ms in the linear AEC module. The computational time per frame of the GCCRN model is 2.3 ms. And thus the total computational time to process a frame is 2.62 ms, which is much less than the stride time 10 ms.

6. Conclusions

This paper describes a two-stage acoustic echo cancellation and suppression framework for the INTERSPEECH2021 AEC Challenge. We use four parallel partitioned block frequency domain adaptive filters to cancel the linear echo components with a large dynamic time delay in the blind test dataset. The error signal with minimum energy and the corresponding reference signal are chosen as the input for the second stage. In the second stage, a gate complex convolutional recurrent neural network is trained to further suppress the residual echo components, the late reverberation and the environmental noise. The compressed complex spectrum of the error signal and that of the reference signal are used as the input features, where these complex spectra are regenerated with the compressed magnitude and the phase of the error signal. Experimental results show that the proposed framework is quite robust to the blind test dataset, which achieves a good result in the P.808 evaluation.

7. Acknowledgements

This work was support by the National Science Foundation of China under Grant No. 61801468, and IACAS Young Elite Researcher Project under Grant No. QNYC201813.

8. References

[1] B. Farhang, *Adaptive filters: theory and applications*. John Wiley & Sons, 2013.

[1] https://github.com/microsoft/AEC-Challenge/tree/main/DECMOS

[2] A. Munjal, V. Aggarwal, and G. Singh, “Rls algorithm for acoustic echo cancellation,” *COIT, March*, vol. 29, 2008.

[3] K. Eneman and M. Moonen, “Iterated partitioned block frequency-domain adaptive filtering for acoustic echo cancellation,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 2, pp. 143–158, 2003.

[4] M. M. Halimeh and W. Kellermann, “Efficient multichannel nonlinear acoustic echo cancellation based on a cooperative strategy,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 461–465.

[5] L. Ma, H. Huang, P. Zhao, and et al, “Acoustic echo cancellation by combining adaptive digital filter and recurrent neural network,” *arXiv preprint arXiv:2005.09237*, 2020.

[6] L. W. Nils and T. M. Bernd, “Acoustic echo cancellation with the dual-signal transformation lstm network,” 2020.

[7] J. M. Valin, S. Tenneti, K. Helwani, and et al, “Low-complexity, real-time joint neural echo control and speech enhancement based on percepnet,” 2021.

[8] R. H. Peng, L. J. Cheng, C. S. Zheng, and X. D. Li, “Icassp 2021 acoustic echo cancellation challenge: Integrated adaptive echo cancellation with time alignment and deep learning-based residual echo plus noise suppression,” in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1–5.

[9] K. Sridhar, R. Cutler, A. Saabas, and et al, “Icassp 2021 acoustic echo cancellation challenge: Datasets and testing framework,” *arXiv preprint arXiv:2009.04972*, 2020.

[10] B. Sebastian, G. Hannes, C. A. Reddy, and I. Tashev, “Towards efficient models for real-time deep noise suppression,” *arXiv preprint arXiv:2101.09249*, 2021.

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

[12] A. D. Li, C. S. Zheng, R. H. Peng, and X. D. Li, “On the importance of power compression and phase estimation in monaural speech dereverberation,” *JASA Express Letters*, vol. 1, no. 1, pp. 1–7, 2021.

[13] T. P. M. L. S. S. M. P. H. G. S. B. K. S. R. A. S. S. Ross Cutler, Ando Saabas, “Interspeech 2021 acoustic echo cancellation challenge: Datasets and testing framework,” in *INTERSPEECH 2021*.

[14] K. Park, “Korean single speaker speech dataset,” https://www.kaggle.com/bryanpark/korean-single-speaker-speech-dataset.

[15] DiDi, “Gaia open dataset,” https://outreach.didichuxing.com/research/opendata/.

[16] C. K. Reddy, E. Beyrami, H. Dubey, and et al, “The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework,” 2020.

[17] Openslr, “Mediaspeech,” http://www.openslr.org/108.

[18] U. of Tokyo, “Japanese speech corpus of saruwatari-lab., university of tokyo,” https://sites.google.com/site/shinnosuketakamichi/publication/jsut.

[19] Openslr, “Thorsten mller,” http://www.openslr.org/95.

[20] K. Tom, P. Vijayaditya, P. Daniel, and et al, “A study on data augmentation of reverberant speech for robust speech recognition,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[21] C. K. Reddy, E. Beyrami, H. Dubey, and et al, “The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework,” *arXiv preprint arXiv:2001.08662*, 2020.

[22] I. Rec, “P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs,” *International Telecommunication Union, CH–Geneva*, 2005.

[23] R. Culter, A. Saabas, T. Parnamaa, and et al, “Interspeech 2021 acoustic echo cancellation challenge,” *arXiv preprint arXiv:2009.04972*, 2020.