Acoustic Echo Cancellation Based on Recurrent Neural Network

Yao-Cheng Tsai [1], Kai-Wen Liang[1], and Pao-Chi Chang[1]
[1]National Central University, Taoyuan, Taiwan
E-mail: yctsai.vaplab@gmail.com, kwistron@gmail.com, pcchang@ce.ncu.edu.tw

Abstract—This work proposes an acoustic echo cancellation method using deep-learning-based speech separation techniques. Traditionally, acoustic echo cancellation (AEC) used a linear adaptive filter to identify the acoustic impulse response between the microphone and the loudspeaker. However, when conventional methods encounter nonlinear conditions, the results of the processing are not good enough. Our practice utilizes the advantages of deep-learning techniques, which are beneficial for nonlinear processings. In the adopted recurrent neural network system, we add single-talk features and assign specific weighting for each element in different from the traditional speech separation. The experimental results show that our method improves the Perceptual evaluation of speech quality (PESQ) of simulated audio, and the Echo return loss enhancement (ERLE) of recorded audio as well.

Keywords—Deep Learning, Acoustic Echo Cancellation, Speech Separation, Recurrent Neural Network

I. INTRODUCTION

The conventional acoustic echo cancellation method uses a linear adaptive filter to identify the acoustic impulse response between the microphone and the loudspeaker. Fig.1 shows the block diagram of our echo cancellation system, where $x(n)$, $h(n)$, $d(n)$, $s(n)$, $v(n)$, $y(n)$, $\hat{s}(n)$, respectively far-end signal, acoustic echo path, echo signal, near-end signal, background noise, microphone signal, resynthesize signal. Fig.2 shows the internal structure diagram of the AEC, which is an acoustic echo cancellation method using a linear adaptive filter. The estimated near-end signal is obtained by subtracting the estimated far-end echo signal from the microphone signal.

II. RELATED BACKGROUND

In 1990, J.-S. Soo et al. proposed a multi-delay block frequency domain adaptive filter (MDF) [1] based on a normalized least mean square (NLMS) algorithm. NLMS algorithm calculated the input signal in the time domain and updated the weights frame by frame, which cost a large amount of time consumption. The MDF algorithm performed calculation in the frequency domain, batch-processed input signals, and uniformly performed weight updating. An implementation of the MDF algorithm is available in Speex [2], which is an open-source algorithm and compared in our experiments. In 2017, D. Yu et al. proposed a permutation-invariant training (PIT) speech separation method [3]. Their approach trained the neural network by minimizing the estimation error between the original speech and the estimated speech. In 2018, Zhang, H et al. proposed a deep learning method for acoustic echo cancellation [4], which also used speech separation techniques for acoustic echo cancellation by mixed-signal and features of far-end echo signals. Our proposed method jointly references the features of the near-end signal and the far-end echo signal at the same time and adjusts the weights of different features. The experimental results show that the ERLE and the PESQ are improved. We train our model through a bidirectional gated recurrent unit (BGRU) [5] [6] to build the neural networks.

The proposed methods will be described in section III. Section IV and V will specify the evaluation criterion and discuss the experiments, respectively. Finally, section VI concludes this paper.

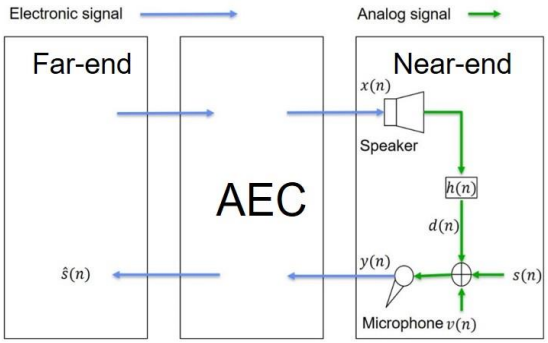

Fig. 1 Block diagram of our echo cancellation system

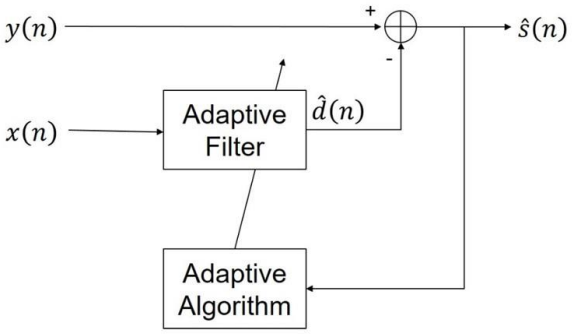

Fig. 2 Internal structure diagram of the AEC

III. PROPOSED METHOD

In this section, we will introduce our proposed method which is based on recurrent neural networks [7]. Section *A* will present the pre-processing of the dataset, and Section *B* illustrates the training stage; finally, Section *C* will describe the testing stage. Fig.3 shows the architecture diagram of the proposed method.

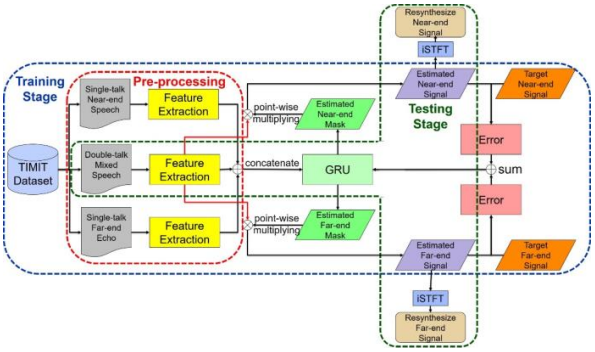

Fig. 3 Architecture diagram of the proposed method

A. *Pre-processing*

In this work, we extract the features of speech data by a 256 points short-time Fourier transform (STFT) with 129frequency bins. The sampling rate is 16kHz, and the frame size is 16ms. Fig.4 illustrates the flow chart of STFT.

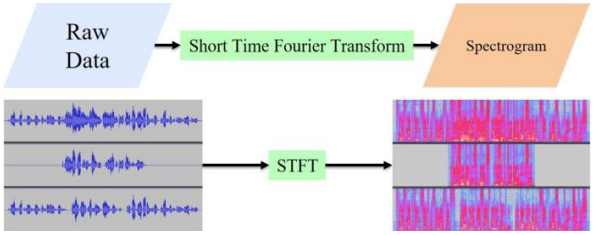

Fig.4 Flow chart for short-time Fourier transform

B. *Training Stage*

We adopt a recurrent neural network structured by BGRU in the proposed method. In different from the bidirectional long short-term memory (BLSTM) network [6][8], the proposed method is less complicated. By adding the single-talk features (near-end signal and far-end echo signal) and adjusting the weights of each element, the results show good performance in our experiments. We use the ideal ratio mask [9] as the speech separation target. Eq (1) and (2) express the mathematical calculation of the masks.

$$IRM_1(t,f) = \sqrt{\frac{{X_1}^2(t,f)}{Y^2(t,f)}} \quad (1)$$

$$IRM_2(t,f) = \sqrt{\frac{{X_2}^2(t,f)}{Y^2(t,f)}} \quad (2)$$

where X_1, X_2 and Y are the target near-end signal, the target far-end echo signal, and the microphone signal. The ideal ratio mask is the ratio of the target signal in the mixed-signal, which is defined as our training target. The ideal ratio mask value is between 0 and 1, so we use ReLU as the activation function to limit it within this range. Both the input layer and the output layer of the BGRU are with129 neurons. There are three hidden layers with 496 neurons for each. The learning rate is 0.0005, and the number of iterations is 60. The loss function used in BGRU is the mean square error (MSE), as expressed in Eq (3).

$$J_X = \frac{1}{T\times F\times S}\sum_{t=1}^{T}\sum_{f=1}^{F}\sum_{s=1}^{S}\left\|\hat{X}_s(t,f) - X_s(t,f)\right\|^2 \quad (3)$$

where T is the number of time frames, F is the number of frequency bins, $\hat{X}_s$ is the estimated speaker's signal, X_s is the target speaker's signal, S is the number of speakers.

C. *Testing Stage*

In the testing stage, the features of the mixed-signal for testing are sent into BGRU to obtain the estimated signal by pointwise multiplying the features with the estimated mask.

IV. EVALUATION

To reduce the time consumption of the training stage, we use GPU to execute the program for acceleration. We use the RTX 2080 GPU to run our Python program, where we use tensorflow and librosa as tools in our neural network and feature extraction, respectively. In this section, we will introduce the performance metrics in Section *A* and the experimental setup in Section *B*.

A. *Performance Metrics*

We use two performance metrics to evaluate the results of our experiments. In the single-talk scenario, we use ERLE to assess the extent of attenuation to our far-end echo. The higher value indicates the better ability of the system to eliminate echo. In the double- talk scenario, we use PESQ [10] to evaluate the correlation between the estimated signal and the original signal. When the calculated value is high, it means the correlation is high, and therefore it can maintain a certain degree of speech quality. The ERLE mathematical calculation can be expressed in (4), where $y(n)$ is the microphone signal, $\hat{s}(n)$ is the resynthesize signal.

$$ERLE = 10log_{10}\left\{\frac{\varepsilon[y^2(n)]}{\varepsilon[\hat{s}^2(n)]}\right\} \quad (4)$$

B. *Experiment Setup*

We adopted the TIMIT dataset [11] in our experiments. The TIMIT dataset has 630 speakers. Each speaker speaks 10 sentences. There is a total of 6,300 sentences. Each sentence is sampled at 16 kHz. First, we will make these sentences form the near-end signal and the far-end signal. We will randomly select three sentences from a random speaker and connect them to form a far-end signal. We will randomly select another speaker's sentence and pad zeros at the front and the back to the same length as the far-end signal to form a near-end signal. In the far-end signal part, the number of speakers used in the training stage is 30 speakers, the number of speakers used in the test stage is six speakers, and the remaining speakers as the near-end signal. We generate mixed speech in two cases, simulation and live recording. In the case of simulation, we use the image method [12] to generate a room impulse response to convolve with the far-end signal to derive the far-end echo signal. The simulation room size is (4, 4, 4) meters, the microphone is placed at the center of the room (2, 2, 2) meters, and the loudspeakers are placed at random places. We use the gaussian noise to generate signal-to-noise ratios (SNR)

between 0 dB and 5 dB in the case of simulation. In the case of recording, we record live in the conference room where the doors and windows are closed. Besides that, we can calculate the relationship between the near-end signal and the far-end echo signal by signal-to-echo ratios (SER). We chose SER for 10 dB, 0 dB, and -10 dB in experiment, respectively. The SNR and SER can be expressed in (5) and (6),

$$SNR = 10log_{10}\left\{\frac{\varepsilon[s^2(n)]}{\varepsilon[v^2(n)]}\right\} \quad (5)$$

$$SER = 10log_{10}\left\{\frac{\varepsilon[s^2(n)]}{\varepsilon[d^2(n)]}\right\} \quad (6)$$

where $s(n)$ is the near-end signal, $v(n)$ is the background noise, $d(n)$ is the far-end echo signal. In the case of simulation, the far-end signal convolves with a room impulse response to derive the far-end echo signal. The far-end echo signal mathematical calculation can be expressed in (7),

$$d(n) = x(n) * h(n) \quad (7)$$

where $d(n)$ is the far-end echo signal, $x(n)$ is the far-end signal, $h(n)$ is the room impulse response. The microphone signal mathematical calculation can be expressed in (8),

$$y(n) = d(n) + s(n) + v(n) \quad (8)$$

where $d(n)$ is the far-end echo signal, $s(n)$ is the near-end signal, $v(n)$ is the background noise.

V. Experiment Results

The proposed method will be compared with the MDF algorithm and the PIT speech separation method. The experiments conduct with three different cases, high SER, medium SER, and low SER, in both of the simulation and recording data for comparisons.

The results of ERLE and PESQ in high SER situations are listed in Table 1 and Table 2. In the case of simulation, the ERLE is the highest in BLSTM, but in the case of recording, the training model of the neural networks with the near-end signal feature obtains the most top result of all methods. In the case of simulation, the obtained PESQ is the highest when we add the near-end signal feature and adjust the weights of the features for training. The experimental results show that the ERLE and PESQ can be conditionally improved by our proposed method.

Table 1 High SER (10 dB) of ERLE

Performance Metrics: ERLE (dB)	Simulation	Recording
Speex [2]	19.84	9.68
BLSTM [3] (mixed speech feature)	42.29	37.98
Proposed BGRU (mixed speech feature)	38.16	38.82
Proposed BGRU (mixed + near-end feature)	38.79	43.52
Proposed BGRU (mixed*10% + near*90% feature)	32.04	21.06
Proposed BGRU (mixed*90% + near*10% feature)	36.67	20.57

Table 2 High SER (10 dB) of PESQ

Performance Metrics: PESQ	Simulation	Recording
Speex [2]	3.35	2.95
BLSTM [3] (mixed speech feature)	2.37	2.15
Proposed BGRU (mixed speech feature)	2.47	2.21
Proposed BGRU (mixed + near-end feature)	2.5	1.48
Proposed BGRU (mixed*10% + near*90% feature)	2.44	2.03
Proposed BGRU (mixed*90% + near*10% feature)	2.65	1.88

Table 3 and Table 4 list the results of ERLE and PESQ in a medium SER situation. The overall results are not as high as in the High SER situation. Similarly, the ERLE in the case of simulation in BLSTM is the highest, while our method performs the best in the case of the recording. Similar results as in the High SER situation can be found in the case of simulation.

Table 3 Medium SER (0 dB) of ERLE

Performance Metrics: ERLE (dB)	Simulation	Recording
Speex [2]	19.48	9.84
BLSTM [3] (mixed speech feature)	36.47	40.03
Proposed BGRU (mixed speech feature)	32.02	36.98
Proposed BGRU (mixed + near-end feature)	32.56	44.83
Proposed BGRU (mixed*10% + near*90% feature)	18.41	17.48
Proposed BGRU (mixed*90% + near*10% feature)	29.91	11.41

Table 4 Medium SER (10 dB) of PESQ

Performance Metrics: PESQ	Simulation	Recording
Speex [2]	3.14	2.4
BLSTM [3] (mixed speech feature	1.97	0.87
Proposed BGRU (mixed speech feature)	1.97	0.98
Proposed BGRU (mixed + near-end feature)	2	0.85
Proposed BGRU (mixed*10% + near*90% feature)	2.03	0.98
Proposed BGRU (mixed*90% + near*10% feature)	2.22	0.92

Finally, it can be seen in Table 5 and Table6, even in the Low SER situation, our method can improve the performance in certain conditions.

Traditional speech separation uses only mixed-signal features, while our proposed method adds a single-talk feature and adjusts the weight of each element. The experimental results show that in the case of recording, after adding the single-talk features, the ERLE reaches the highest of all methods. In the case of simulation, after adding the single-talk features and adjusting the weights of elements, the PESQ also achieves the highest in the deep learning methods.

Table 5 Low SER (-10 dB) of ERLE

Performance Metrics: ERLE (dB)	Simulation	Recording
Speex [2]	18.77	12.48
BLSTM [3] (mixed speech feature)	33.51	35.15
Proposed BGRU (mixed speech feature)	28.63	29.19
Proposed BGRU (mixed + near-end feature)	28.43	35.98
Proposed BGRU (mixed*10% + near*90% feature)	13.49	13.63
Proposed BGRU (mixed*90% + near*10% feature)	22.9	21.98

Table 6 Low SER (-10 dB) of PESQ

Performance Metrics: PESQ	Simulation	Recording
Speex [2]	2.87	2.08
BLSTM [3] (mixed speech feature)	1.46	0.86
Proposed BGRU (mixed speech feature)	1.41	0.78
Proposed BGRU (mixed + near-end feature)	1.43	0.92
Proposed BGRU (mixed*10% + near*90% feature)	1.66	0.76
Proposed BGRU (mixed*90% + near*10% feature)	1.71	0.78

VI. CONCLUSIONS

This work proposes a bidirectional gated recurrent unit (BGRU) based acoustic echo cancellation, which adds near-end features. The experimental results show that the performance can be improved by adding single-talk features and adjusting the weights of elements. In the case of simulation, our proposed method is 0.2-0.3 better than BLSTM [3]. In the case of recording, the proposed method is 20-30 dB better than Speex [2]. The proposed method could offer improved PESQ in the simulation cases while the performance on ERLE decreases lightly. In the future, we will add various recording data in the training stage to enhance the system model for different environments.

REFERENCES

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

[2] J.-M. Valin, Speex: A Free Codec For Free Speech, Available: http://www.speex.org/

[3] M. Kolbak, D. Yu, Z.-H. Tan, and J. Jensen, "Multi-talker speech separation with utternance-level permutation invariant training of deep recurrent neural networks" *IEEE/ACM Trans. Audio Speech Lang. Proc*., vol. 25, pp. 1901-1913, 2017.

[4] Zhang, Hao, and DeLiang Wang. "Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios," Training 161.2 (2018): 322.

[5] Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." arXiv preprint arXiv:1412.3555 (2014).

[6] Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." *IEEE Transactions on Signal Processing* 45.11 (1997): 2673-2681.

[7] A. Graves, A. Mohamed, and G. Hinton, "Speech recognitionwith deep recurrent neural networks," in *ICASSP*, 2013, pp.6645–6649

[8] S. Hochreiter and J. Schimidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997

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

[10] 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 *ICASSP*, 2001, pp. 749–752.

[11] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in Speech Input/Output Assessment and Speech Databases, 1989.

[12] J. B. Allen and D. 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.