Acoustic echo cancellation based on two-stage BLSTM

Zhiwei Niu,[1] Shifeng Ou,[1] Peng Song,[2] and Ying Gao[1,✉]

[1]School of Physics and Electronic Information, Yantai University, Yantai, China

[2]School of Computer and Control Engineering, Yantai University, Yantai, China

✉ E-mail: claragaoying@126.com

Acoustic echo cancellation (AEC) methods aim to suppress the acoustic coupling for hands-free speech communication. Traditional AEC works by identifying the acoustic impulse response using adaptive algorithms. With recent research advances, deep learning has become an attractive choice for AEC. This paper introduces a two-stage bidirectional long short term memory (TS-BLSTM) framework, incorporating multi-head self-attention mechanisms after each BLSTM block. This is aimed at better capturing contextual information and further enhancing ability of the model to handle complex acoustic scenarios. The BLSTM blocks are utilized to aggregate magnitude spectrum information, modelling both time and frequency dependencies. Additionally, dilation convolution is introduced to broaden the range of information in each convolution output. The magnitude decoder estimates a mask for the input, resulting in the generation of an estimated magnitude spectrum for near-end speech. Experimental results indicate that the proposed method achieves promising outcomes.

Introduction: Acoustic echo cancellation (AEC) plays an essential part in hands-free VoIP speech communication and video conferencing systems. Acoustic echo arises in a full-duplex voice communication system when a near-end microphone picks up audio signals from a near-end loudspeaker and sends it back to a far-end participant such that the far-end user receives a modified version of his/her voice [1, 2]. AEC aims to remove the echo from the microphone signal while leaving the near-end speech least distorted.

Conventional methods address this problem by estimating the acoustic path with an adaptive filter. Several adaptive algorithms have been proposed in the literature [3]. Among them the normalized least mean square (NLMS) and affine projection (AP) algorithm family [4, 5] is most widely used. However, when faced with double-talk situations, the presence of a near-end speech signal severely degrades the convergence of adaptive algorithms and may cause them to diverge. To address the double-talk issue, the adaptive filter can be linked with double-talk detectors (DTD) to halt filter adaptation during instances of double-talk [6]. Alternatively, the adaptive filter itself can be designed to withstand double-talk by incorporating robust criteria, as suggested in [7].

Moreover, many studies in the literature model the echo path as a linear system. However, due to the limitations of components such as power amplifiers and loudspeakers, a nonlinear distortion may be introduced to the far-end signal in the practical scenario of AEC. To overcome this problem, several nonlinear models such as the Volterra model, the Hammerstein model, and functional link adaptive filters have been utilized [8–10]. Despite such a lot of works, the adaptive filtering approach still has not shown satisfactory results in various environments.

With the advancement of deep learning, several speech processing tasks, including speech recognition [11, 12], speech enhancement [13], and speech separation [14], have been successfully undertaken using neural networks. Acoustic echo cancellation has also been addressed through various innovative solutions. For instance, Seo et al. [15] introduced a stacked deep neural network (DNN) model, incorporating one component for noise suppression and another for acoustic echo suppression in a sequential manner to handle both acoustic echo and background noise. In a different approach, a bidirectional long short-term memory (BLSTM) model was employed by [16] to predict the ideal ratio mask from microphone signals and subsequently use it for the resynthesizes of proximal speech. A recurrent neural network (RNN) with multi-task learning was proposed by [17], addressing the auxiliary task of estimating echoes to enhance the primary task of estimating near-end speech.

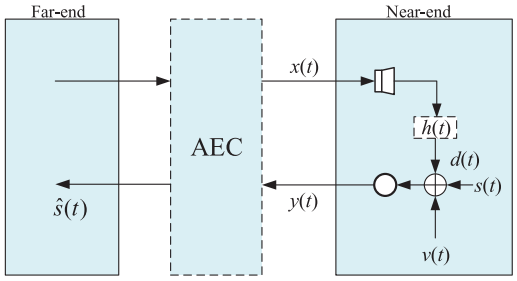

Fig. 1 *Diagram of acoustic echo scenario*

Notably, Kim and Chang [18] introduced an attentional wave-U-network for AEC, featuring an auxiliary encoder for the extraction of features from distal speech. In the context of real-time AEC, Zhang et al. [19] implemented a complex neural network to more effectively model critical phase information, alongside a frequency-time-LSTM (F-T-LSTM) scanning both frequency and time axes for improved temporal modelling, using complex Conv2d layers and complex transposed-Conv2d layers as encoder and decoder respectively to model the complex spectra from both far-end and near-end signal. Halimeh et al.[20] introduced a novel approach to noise-robust acoustic echo cancellation by employing a complex-valued DNN for postfiltering. This integrated approach, blending adaptive filtering and deep learning technologies, offers a wealth of insights for advancing echo cancellation within various applications. Han et al.[21] proposed a novel speaker- and phone-aware dual-path convolutional transformer network (DPCT-Net) for AEC, the speaker- and phone-aware DPCTNet consists of a CPC network to extract representations that contain phonetic and speaker identities information and a complex dual-path convolutional transformer network which is encoder–decoder structure to learn and reconstruct near-end speech. Furthermore, a conditional generative adversarial network (cGAN) based AEC system was presented in [22]. The generator of the proposed cGAN framework is composed of a U-Net model able to synthesize the echo-free signal. The synthesized signal, conditioned by the estimated echo signal, is the input to the discriminator. The discriminator aims to refine the synthesized signals to convert them as realistic as possible.

In this paper, we build upon the foundation of BLSTM, leveraging its bidirectional reading capability to focus on learning speech sequence information. Breaking through the structure of the network model itself, we construct a two-stage BLSTM network, aggregating time and frequency dependencies through amplitude spectrum modelling. We employ multi-head self-attention (MHSA) mechanism to guide the model in establishing mapping relationships. In both the encoder and decoder sections, dilation convolution is introduced to increase the receptive field without adding parameters, enabling the model to better capture long-distance contextual information. Experimental results indicate that the proposed method significantly improves the effectiveness of echo cancellation.

Problem formulation: In the AEC application, as shown in Figure 1, the microphone signal $y(t)$ is composed of acoustic echo $d(t)$, near-end signal $s(t)$, and background noise $v(t)$ as follows:

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

where $d(t)$ is the acoustic echo, which can also be nonlinearly distorted by the loudspeaker, is a modified version of the far-end signal by a room impulse response (RIR). The purpose of the AEC is to estimate the near-end signal $s(t)$ from the microphone signal $y(t)$ by suppressing the acoustic echo $d(t)$.

As illustrated in Figure 2, our network consists of three modules: magnitude encoder, magnitude decoder and TS-BLSTM blocks. For the microphone speech waveform $y \in \mathbb{R}^{B\times L}$ and the far-end speech waveform $x \in \mathbb{R}^{B\times L}$, the complex spectrogram X_r, X_i, Y_r, and Y_i are obtained by short-time Fourier transform (STFT), where B and L denote batch size

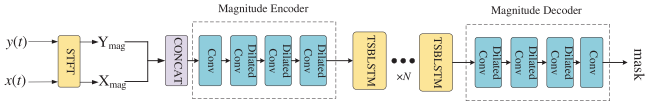

Fig. 2 *Proposed framework to perform acoustic echo cancellation*

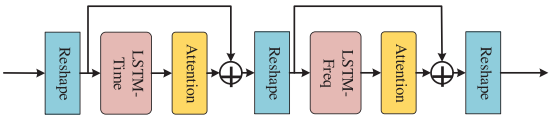

Fig. 3 *Two-stage BLSTM (TS-BLSTM)*

and number of speech sampling points, respectively. Extract the amplitudes of X and Y, denoted X_{mag} and Y_{mag}, respectively, and combine the microphone signal and the far-end signal into two channels $\mathbb{R}^{B\times C\times T\times F}$, with C denotes channels. The amplitudes of X and Y can be obtained by

$$\begin{cases} X_{\text{mag}} = \sqrt{X_r^2 + X_i^2} \\ Y_{\text{mag}} = \sqrt{Y_r^2 + Y_i^2} \end{cases} \tag{2}$$

All audio signals are sampled at 16 kHz and the preprocessing is performed identically for reference far-end and microphone signals. The input features to the network are log power spectra computed with a squared root Hann window. We adopt a frame-by-frame approach to process audio data. We divide the entire audio signal into small time segments, and each frame is processed individually. This method allows us to perform real-time processing of audio data while preserving the temporal correlation of the audio data. The method proposed in this section incurs a delay of 0.0769 s for generating 1 s of speech. Generally, when the delay is less than 0.1 s, it can be considered as real-time processing. Therefore, the proposed method supports real-time processing.

Phase information is typically more challenging to handle than amplitude information because it is random without fixed patterns. BLSTM is better at capturing long-term time dependencies and dynamic patterns in sequences when dealing with temporal data, but it faces the risk of information loss when processing phase information. The structure of BLSTM networks is more complex, requiring bidirectional recurrent neurons, leading to higher computational complexity.

Magnitude encoder: The magnitude encoder is composed of a convolution block and a dilated DenseNet [23]. The convolution block comprises a two-dimensional convolution layer, an instance normalization [24] and a PReLU activation [25]. The convolution kernel size of the two-dimensional convolution is set to (1,1), and the step size is set to (1,1). The shape of the tensor remains unchanged, only the number of channels is changed, and low-dimensional channel 2 is convolved to high-dimensional 64 channels for feature extraction in higher dimensions. The dilated DenseNet contains three convolution blocks with dense connections, the dilation factors of each block are set to 1, 2, 4. The dense connections can aggregate all previous feature maps to extract different feature levels. As for the dilated convolutions, they serve to increase the receptive field effectively while preserving the kernels and layers count.

TS-BLSTM: LSTM [26] networks can store information in their memory layer, but are limited in their ability to compute in only one direction for signal data, thereby failing to fully capture the temporal correlations in the data. BLSTM networks offer a better solution to this problem by employing a forward LSTM to retain preceding hidden features and a backward LSTM to retain subsequent hidden features. As speech signals exhibit strong both forward and backward correlations, BLSTM tends to achieve better performance compared to LSTM. The combination of the gate mechanism and memory cells in the BLSTM model exhibits significant effectiveness in capturing speech signals. The introduction of dilated convolution further enhances the model's ability to capture long-term dependencies. Simultaneously, the multi-head self-attention mechanism provides the model with the capability to adaptively focus on contextual information at different scales. These design elements collectively contribute to the model's superior performance in handling nonlinear distortions in the context of processing speech signals. As shown in the Figure 3, given a feature map $M \in \mathbb{R}^{B\times C\times T\times F}$, the input feature map M is first reshaped to $M^T \in \mathbb{R}^{BF\times T\times C}$ to capture the time dependency in the first BLSTM block. Then the output is element-wise added with the input M^T (residual connection) and reshaped to a new feature map $M^F \in \mathbb{R}^{BT\times F\times C}$. The second BLSTM thus captures the frequency dependency. After the residual connection, the final output is reshaped back to the input size. After each BLSTM block, we employed MHSA mechanism [27] with four heads. The MHSA mechanism offers distinct advantages in echo cancellation by enabling the model to simultaneously focus on different aspects of the input sequence. This capability facilitates the comprehensive learning of features, including both temporal and spectral dependencies, crucial for understanding the complex interplay of time and frequency components in acoustic signals. The mechanism enhances robustness by adapting to variations in the input signal, establishing intricate mapping relationships, and effectively modelling long distance contextual information. The organic integration of BLSTM blocks and MHSA achieved a more comprehensive joint time-frequency modelling. This enhancement improves the model's ability to model complex audio scenes, bringing about more accurate performance for the task of echo cancellation.

Table 1. Configuration of our proposed method

Layer name	Input size	Hyperparameters	Output size
Conv2d	$2 \times T \times 161$	(1,1),(1,1)	$64 \times T \times 161$
Dilated Conv	$64 \times T \times 161$	$(2, 3), (1, 1), d = 1$	$64 \times T \times 161$
Dilated Conv	$128 \times T \times 161$	$(2, 3), (1, 1), d = 2$	$64 \times T \times 161$
Dilated Conv	$192 \times T \times 161$	$(2, 3), (1, 1), d = 4$	$64 \times T \times 161$
Reshape	$64 \times T \times 161$	–	$161 \times T \times 64$
LSTM-time	$161 \times T \times 64$	128	$161 \times T \times 64$
Reshape	$161 \times T \times 64$	–	$T \times 161 \times 64$
LSTM-freq	$T \times 161 \times 64$	128	$T \times 161 \times 64$
Reshape	$T \times 161 \times 64$	–	$64 \times T \times 161$
Dilated Conv	$64 \times T \times 161$	$(2, 3), (1, 1), d = 1$	$64 \times T \times 161$
Dilated Conv	$128 \times T \times 161$	$(2, 3), (1, 1), d = 2$	$64 \times T \times 161$
Dilated Conv	$192 \times T \times 161$	$(2, 3), (1, 1), d = 4$	$64 \times T \times 161$
Conv2d	$64 \times T \times 161$	(1,1),(1,1)	$1 \times T \times 161$

Magnitude decoder: The magnitude decoder extracts the output from four TS-BLSTM blocks in a decoupled manner, i.e. mask decoder. The mask decoder aims to predict a mask that will be element-wise multiplied with the input magnitude. The magnitude encoder is composed of a dilated DenseNet and a convolution block. The convolution layer is used to compress the channel from 64 to 1, and the sigmoid activation function is used to get the final predicted mask. Multiply the mask obtained from the output by the amplitude spectrum of the microphone signal to obtain the estimated amplitude spectrum of the near-end speech. Obtain the estimated composite spectrum from Equation (3), and further calculate the inverse STFT (ISTFT) to obtain the estimated time-domain signal. Our experimental parameters are shown in Table 1, where d represents the dilation factor.

$$\begin{cases} \hat{S}_r = \hat{S} \cdot \cos\theta \\ \hat{S}_i = \hat{S} \cdot \sin\theta \end{cases} \tag{3}$$

Loss function: To partially avoid the compensation effect between the magnitude and RI constraints in [28] and [29], we use the linear combination of magnitude and complex loss:

$$\mathcal{L} = \frac{\mathcal{L}_{\text{mag}} + \mathcal{L}_{ri}}{2} \tag{4}$$

$$\mathcal{L}_{\text{mag}} = \mathbb{E}_{S_{\text{mag}},\hat{S}_{\text{mag}}}\left[\left\|S_{\text{mag}} - \hat{S}_{\text{mag}}\right\|^2\right] \tag{5}$$

$$\mathcal{L}_{ri} = \mathbb{E}_{S_r,\hat{S}_r}\left[\left\|S_r - \hat{S}_r\right\|^2\right] + \mathbb{E}_{S_i,\hat{S}_i}\left[\left\|S_i - \hat{S}_i\right\|^2\right] \tag{6}$$

For the weight selection, this paper adopts equal weight allocation for both amplitude loss and composite loss to reflect their equal importance in the overall optimization process. This choice of weights is based on a profound understanding of acoustic signals, taking into account not only the mathematical characteristics of the loss functions but also the perceptual attributes of speech signals. Consequently, it enhances ability of the model to model sound details effectively.

We used TIMIT dataset [30] to evaluate AEC performance. From 630 speakers of TIMIT, 100 pairs of speakers (40 male-female, 30 male-male, 30 female-female) are randomly chosen to be used as the far-end and near-end speakers. Each speaker sampled 10 sounds at 16 kHz, randomly selected a speaker is intercepted as a 6 s, and those less than 6 s are expanded to 6s by repetitive stacking. Then, the near-end signals are truncated to 2 s. If it is less than 2 s, zero data is added after it, and then the starting point of each signal is filled with 4 s of zero data. Seven utterances of near-end speakers are used to generate 3500 training mixtures where each near-end signal is mixed with five different far-end signal. From the remaining 430 speakers, we randomly picked another 100 pairs of speakers as the far-end and near-end speakers. We followed the same procedure as described above and obtained 300 validation sets and test sets.

The following processes were applied to the far-end signal to model the nonlinear acoustic path as in [31]. For the nonlinear model of acoustic path, we first applied the hard clipping to simulate the power amplifier of loudspeaker ($x_{\max}$ is set to 80% of the maximum volume of input signal):

$$x_{\text{clip}}(t) = \begin{cases} -x_{\max} & x(t) < -x_{\max} \\ x(t) & |x(t)| \le x_{\max} \\ x_{\max} & x(t) > x_{\max} \end{cases} \tag{7}$$

Then, to simulate the loudspeaker distortion, we applied the following sigmoidal function:

$$x_{nl}(t) = 4\left(\frac{2}{1+\exp(-a \cdot b(t))} - 1\right) \tag{8}$$

where,

$$b(t) = 1.5 \times x_{\text{clip}}(t) - 0.3 \times x_{\text{clip}}^2(t) \tag{9}$$

If $b(t) > 0$, then the slope a is set to 4, otherwise it is set to 0.5. Finally, the output of sigmoidal function is convolved with a randomly chosen RIR $g(t)$ in order to simulate the acoustic transmission of far-end signal in the room:

$$d_{nl}(t) = x_{nl}(t) * g(t) \tag{10}$$

where * indicates convolution.

Experiment setting: The length of RIRs is set to 512, the simulation room size is 4 m × 4 m × 3 m, and a microphone is fixed at the location of [2 2 1.5] m. A loudspeaker is placed at seven random places with 1.5 m distance from the microphone. The RIRs are generated using image method [32] with reverberation time (T60) of 350 ms. From 7 RIRs, we used the first six RIRs to generate training data and the last one is used to generate testing data. We also modelled a linear acoustic path by only convolving the far-end signal with RIR to generate the echo signal, clipping and loudspeaker distortion are not applied for this model:

$$d_l(t) = x(t) * g(t) \tag{11}$$

For training mixtures, we generated the microphone signals at signal to echo ratio (SER) level randomly chosen from [0, 3.5, 7] dB by mixing the near-end speech signal and echo signal. For test mixtures, we generated the microphone signals at three different SER levels (0 dB, 3.5 dB, and 7 dB). The SER level is calculated on the double-talk period as:

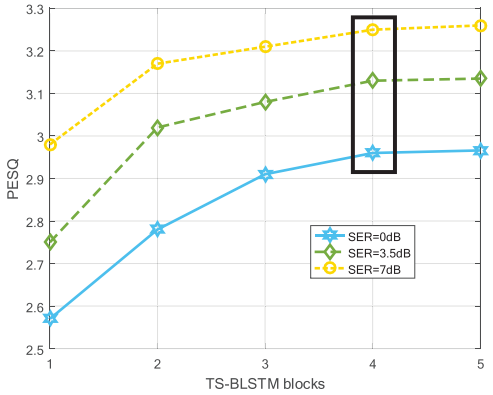

Fig. 4 *Influence of TS-BLSTM blocks on objective scores*

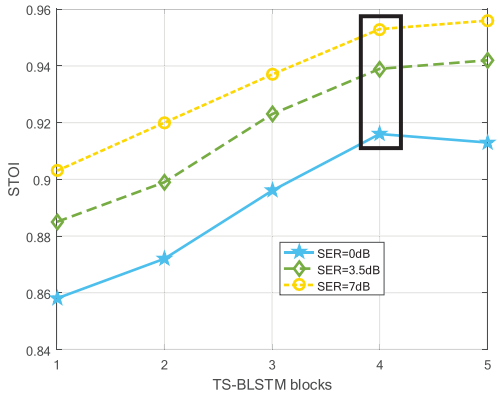

Fig. 5 *Influence of TS-BLSTM blocks on objective scores*

$$\text{SER(dB)} = 10\log_{10}\frac{E\{s^2(t)\}}{E\{d^2(t)\}} \tag{12}$$

Performance metrics: Here, we use three metrics to evaluate the performance of AEC: perceptual evaluation of speech quality (PESQ) [33] and short time objective intelligibility (STOI) [34] for dual ended speech, and echo return loss enhancement (ERLE) [35] for single-talk speech. PESQ is the most commonly used indicator for evaluating speech quality, which is obtained by comparing estimated near-end speech with real near-end speech. The PESQ score ranges from −0.5 to 4.5, with higher scores indicating better quality. STOI is an important indicator for measuring speech intelligibility. Its value range is quantified between 0 and 1, representing the percentage of words correctly understood. A value of 1 indicates that the speech can be fully understood. ERLE is commonly used to evaluate the echo reduction achieved by systems without near-end signals. ERLE is defined as:

$$\text{ERLE(dB)} = 10\log_{10}\frac{E\{y^2(t)\}}{E\{\hat{s}^2(t)\}} \tag{13}$$

System setup: For software, we use PyTorch 1.11.0, Python 3.8, Cuda 11.3, and ubuntu 20.04. For hardware, we used a 24 vCPU AMD EPYC 7642 48 Core Processor and two NVIDIA RTX3090 24GB GPUs. We used Adam [36] optimizer for all networks and the batch size is set to 12. The learning rate is set to 0.0005. The number of training epochs is set to 100. If there is no loss reduction in the validation set for three consecutive epochs, the learning rate needs to be halved.

Experimental results: To justify the model design, we conducted ablation experiments on the number of blocks of TS-BLSTM. As shown in Figures 4 and 5, the scores of PESQ and STOI were effectively improved as the number of blocks was increased, but when we increased the number of blocks from four to five, no significant improvement in performance was observed for PESQ. For STOI, the performance of 5 blocks

Table 2. The influence of different RT60 values

RT60	RIR		PESQ	STOI	ERLE
0.25s	512	None	1.66	0.852	–
		Proposed	3.15	0.941	60.03
0.35s	512	None	1.64	0.850	–
		Proposed	3.11	0.936	62.79
0.35s	1024	None	1.61	0.839	–
		Proposed	3.07	0.927	61.70
0.35s	5600	None	1.61	0.838	–
		Proposed	3.07	0.926	61.68

Table 3. Linear model of acoustic path

Metrics	Method	−3.5 dB	0 dB	3.5 dB	7 dB
PESQ	None	1.44	1.56	1.65	1.71
	BLSTM	2.35	2.49	2.67	2.79
	cGAN	2.41	2.56	2.84	3.07
	TS-BLSTM	2.67	2.85	3.01	3.14
	TS-BLSTM-A	2.82	2.96	3.13	3.25
STOI	None	0.763	0.798	0.854	0.898
	BLSTM	0.824	0.859	0.880	0.901
	cGAN	0.831	0.855	0.888	0.905
	TS-BLSTM	0.877	0.910	0.932	0.948
	TS-BLSTM-A	0.890	0.916	0.939	0.953
ERLE	None	–	–	–	–
	BLSTM	54.01	53.66	51.09	49.37
	cGAN	59.16	58.70	56.24	55.87
	TS-BLSTM	64.08	63.81	62.76	61.68
	TS-BLSTM-A	64.29	63.85	62.82	61.71

at 0 dB SER is even slightly lower than that of 4 blocks. Although the scores of STOI as a whole show an increasing trend, we need to consider the computational volume and the size of the model, and we decided to use four blocks to design our model.

Considering the influence of different RT60 values on the experiments, we conducted comparative experiments with RT60 values of 250 and 350 ms. From the Table 2, one can see that as the reverberation time decreases, both PESQ and STOI have a certain degree of improvement, but ERLE has a certain loss. With an increase in RIR, it can be inferred that there is a small degree of loss in PESQ, STOI, and ERLE. At the same time, we can observe that when the RIR is 1024 and 5600, the experimental results remain consistent.

We first evaluated the proposed method using a linear model of the acoustic path. In order to verify the superiority of the proposed method, two baseline algorithms are used in this paper: BLSTM [17] and cGAN [22]. We set a signal-to-noise ratio (SNR) of 10 dB and conduct comparison tests at different SER, respectively. PESQ, STOI, and ERLE are used as evaluation metrics for the experiments so that the performance of different algorithms can be visualized. The results of the testset under various SER conditions are shown in Table 3.

As can be seen from the Table 3, for the testset data, our proposed TS-BLSTM and TS-BLSTM with the MHSA (TS-BLSTM-A) methods achieved superior results to the other two reference algorithms, regardless of the SER. The experimental results demonstrate the superiority of our proposed TS- BLSTM to carry out feature extraction in two proposed TS- BLSTM to carry out feature extraction in two stages by carrying out features at time and frequency separately. Based on the TS-BLSTM, we added the MHSA mechanism and used its Q, K, and V matrices to perform correlation calculations on the whole input

Table 4. Nolinear model of acoustic path

Metrics	Method	−3.5 dB	0 dB	3.5 dB	7 dB
PESQ	None	1.36	1.48	1.57	1.65
	BLSTM	2.17	2.26	2.49	2.60
	cGAN	2.54	2.63	2.89	3.04
	TS-BLSTM	2.51	2.62	2.81	2.94
	TS-BLSTM-A	2.63	2.74	2.98	3.05
STOI	None	0.755	0.772	0.835	0.880
	BLSTM	0.850	0.870	0.901	0.919
	cGAN	0.858	0.869	0.897	0.899
	TS-BLSTM	0.874	0.889	0.913	0.921
	TS-BLSTM-A	0.879	0.899	0.927	0.940
ERLE	None	–	–	–	–
	BLSTM	51.60	51.33	49.50	47.21
	cGAN	57.73	57.46	55.91	53.95
	TS-BLSTM	59.79	59.57	57.01	54.72
	TS-BLSTM-A	59.91	61.68	58.75	56.54

Table 5. The model size and computational complexity

Model	Params(M)	FLOPs(G)
BLSTM	8.09	0.002
cGAN	1.55	6.76
TS-BLSTM	2.03	1.39
TS-BLSTM-A	2.30	3.864

sequence at once, obtaining a local-to-global correspondence, thus the algorithm performance was further improved.

The combination of the gate mechanism and memory cells in the BLSTM model exhibits significant effectiveness in capturing the characteristic of speech signals. The introduction of dilated convolution further enhances the model's ability to capture long-term dependencies. Moreover, the multi-head self-attention mechanism provides the model with the capability to adaptively focus on contextual information at different scales. These design elements collectively contribute to the model's superior performance in handling nonlinear distortions in the context of processing speech signals. To further study the impact of a nonlinear model of the acoustic path on our scheme, Table 4 presents the objective test results when the acoustic path is nonlinear, which might be caused by power amplifier limiting and loudspeaker distortion. Compared the proposed method with other reference methods, the results showed that TS-BLSTM exhibits slightly lower performance than cGAN in PESQ, but TS-BLSTM-A still outperforms the other algorithms in all three metrics. Thus, the proposed method is experimentally proven to have good echo cancellation effect.

From the Table 5, it can be seen that the proposed model is slightly larger than the cGAN model and smaller than the BLSTM model. Meanwhile, by measuring the model complexity using floating point operations (FLOPs), we can conclude that the complexity of the proposed model is higher than that of BLSTM model and lower than that of cGAN model.

Conclusion: In that study, we proposed a novel acoustic echo cancellation algorithm using a two-stage BLSTM to model the time and frequency domains in two blocks. And we incorporated MHSA mechanism after each block to obtain the time and frequency domain dependencies to predict the mask of the near end speech. Also, dilation convolution is added to increase the range of information. We conducted experiments in the presence of linear and nonlinear distortion respectively, and the proposed algorithm yielded better echo cancellation

performance compared to the competing models, thus demonstrating the effectiveness of our approach in echo cancellation.

Author contributions: **Zhiwei Niu**: Conceptualization; data curation; investigation; software; writing—original draft; writing—review and editing. **Shifeng Ou**: Supervision; writing—review and editing. **Peng Song**: Supervision. **Ying Gao**: Data curation; formal analysis; funding acquisition; methodology; project administration; supervision; validation; writing—review and editing.

Acknowledgments: This work was supported in part by Shandong Provincial Natural Science Foundation under Grant ZR2022MF314.

Conflict of interest statement: The authors declare no conflicts of interest.

Data availability statement: The data that support the findings of this study are available on request from the corresponding author.

Received: *7 January 2024* Accepted: *14 March 2024*
doi: 10.1049/ell2.13164

References

1 Jin, L., et al.: Physical-informed neural network for MPC-based trajectory tracking of vehicles with noise considered. IEEE Trans. Intell. Veh. (2024). https://doi.org/10.1109/TIV.2024.3358229
2 Liufu, Y., et al.: ACP-incorporated perturbation-resistant neural dynamics controller for autonomous vehicles. IEEE Trans. Intell. Veh. (2024). https://doi.org/10.1109/TIV.2023.3348632
3 Hänsler, E., Schmidt, G.: *Acoustic echo and noise control: a practical approach*. Wiley, New York (2005)
4 Shin, H.C., Sayed, A.H., Song, W.J.: Variable step-size NLMS and affine projection algorithms. *IEEE Signal Process Lett.* **11**(2), 132–135 (2004)
5 Tyagi, R., Singh, R., Tiwari, R.: The performance study of nlms algorithm for acoustic echo cancellation. In: International Conference on Information, Communication, Instrumentation and Control (ICICIC), pp. 1–5. IEEE, Piscataway, NJ (2017)
6 Hamidia, M., Amrouche, A.: A new robust double-talk detector based on the stockwell transform for acoustic echo cancellation. *Digital Signal Process.* **60**, 99–112 (2017)
7 Kim, J.H., et al.: Delayless individual-weighting-factors sign subband adaptive filter with band-dependent variable step-sizes. *IEEE/ACM Trans. Audio Speech Lang. Process*. **25**(7), 1526–1534 (2017)
8 Guérin, A., Faucon, G., Le Bouquin-Jeannès, R.: Nonlinear acoustic echo cancellation based on Volterra filters. *IEEE Trans. Speech Audio Process.* **11**(6), 672–683 (2003)
9 Hofmann, C., Huemmer, C., Kellermann, W.: Significance-aware Hammerstein group models for nonlinear acoustic echo cancellation. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5934–5938. IEEE, Piscataway, NJ (2014)
10 Comminiello, D., et al.: Functional link adaptive filters for nonlinear acoustic echo cancellation. *IEEE Trans. Audio Speech Lang. Process*. **21**(7), 1502–1512 (2013)
11 Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Mag*. **29**(6), 82–97 (2012)
12 Li, W., Zhang, P., Yan, Y.: TEnet: target speaker extraction network with accumulated speaker embedding for automatic speech recognition. *Electron. Lett.* **55**(14), 816–819 (2019)
13 Zheng, C., et al.: Sixty years of frequency-domain monaural speech enhancement: from traditional to deep learning methods. *Trends Hearing* **2023**, 27 (2023)
14 Weninger, F., et al.: Discriminatively trained recurrent neural networks for single-channel speech separation. In: 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 577–581. IEEE, Piscataway, NJ (2014)
15 Seo, H., Lee, M., Chang, J.H.: Integrated acoustic echo and background noise suppression based on stacked deep neural networks. *Appl. Acoust.* **133**, 194–201 (2018)
16 Zhang, H., Wang, D.: Deep learning for acoustic echo cancellation in noisy and double-talk scenarios. In: Proceedings of Interspeech 2018, pp. 3239–3243. The Ohio State University, Columbus, OH (2018)
17 Fazel, A., El-Khamy, M., Lee, J.: Deep multitask acoustic echo cancellation. In: Proceedings of Interspeech 2019, pp. 4250–4254. IEEE, Piscataway, NJ (2019)
18 Kim, J.H., Chang, J.H.: Attention wave-u-net for acoustic echo cancellation. In: Proceedings of Interspeech 2020, pp. 3969–3973. IEEE, Piscataway, NJ (2020)
19 Zhang, S., et al.: F-T-LSTM based complex network for joint acoustic echo cancellation and speech enhancement. In: Proceedings of Interspeech 2021, pp. 4758–4762. IEEE, Piscataway, NJ (2021)
20 Halimeh, M.M., et al.: Combining adaptive filtering and complex-valued deep postfiltering for acoustic echo cancellation. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 121–125. IEEE, Piscataway, NJ (2021)
21 Han, C., et al.: Speaker- and phone-aware convolutional transformer network for acoustic echo cancellation. In: Proceedings of Interspeech 2022, pp. 2513–2517 (2022)
22 Pastor-Naranjo, F., et al.: Conditional generative adversarial networks for acoustic echo cancellation. In: 2022 30th European Signal Processing Conference (EUSIPCO), pp. 85–89. IEEE, Piscataway, NJ (2022)
23 Pandey, A., Wang, D.: Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6629–6633. IEEE, Piscataway, NJ (2020)
24 Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Instance normalization: the missing ingredient for fast stylization. arXiv:1607.08022 (2016)
25 He, K., et al.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1026–1034. IEEE, Piscataway, NJ (2015)
26 Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
27 Gulati, A., et al.: Conformer: Convolution-augmented transformer for speech recognition. In: Proceeding Interspeech 2020, pp. 5036–5040. IEEE, Piscataway, NJ (2020)
28 Luo, X., et al.: Analysis of trade-offs between magnitude and phase estimation in loss functions for speech denoising and dereverberation. *Speech Commun.* **145**, 71–87 (2022)
29 Wang, Z.Q., Wang, P., Wang, D.: Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR. *IEEE/ACM Trans. Audio, Speech, Language Process.* **28**, 1778–1787 (2020)
30 Lamel, L., Kassel, R., Seneff, S.: *Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus*. Massachussets Institute of Technology, Cambridge, MA (1989)
31 Malik, S., Enzner, G.: State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation. *IEEE Trans. Audio, Speech, Lang. Process.* **20**(7), 2065–2079 (2012)
32 Allen, J., Berkley, D.: Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.* **65**, 943–950 (1979)
33 Rix, A.W., et al.: Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: 2001-2001 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. **2**, pp. 749–752. IEEE, Piscataway, NJ (2001)
34 Taal, C.H., et al.: A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: 2010-2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4214–4217. IEEE, Piscataway, NJ (2010)
35 Collectif: Academic Press Library in Signal Processing: Image, Video Processing and Analysis, Hardware, Audio, Acoustic and Speech Processing. Academic Press, San Diego, CA (2013)
36 Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 (2014)
37 Xu, Y., et al.: An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process Lett*. **21**(1), 65–68 (2013)