IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 41, No. 4, August, 2024, pp. 1823-1834
Journal homepage: http://iieta.org/journals/ts

D^2L^2-Dense LSTM Deep Learning Based Nonlinear Acoustic Echo Cancellation

D. C. Diana[1*], R. Hema[2], M. Jane Carline[1]

[1] Department of Electronics and Communication Engineering, Easwari Engineering College, Chennai 600089, India
[2] Department of Electronics and Communication Engineering, Saveetha Engineering College, Chennai 602105, India

Corresponding Author Email: diana.d@eec.srmrmp.edu.in

Copyright: ©2024 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ts.410414

Received: 17 December 2023
Revised: 13 April 2024
Accepted: 15 June 2024
Available online: 31 August 2024

Keywords:
non-linear acoustic echo cancellation, deep learning, LSTM, spectral magnitude, source separation, short time Fourier transform

ABSTRACT

Speech quality is a crucial concern, as voice communication is a more noteworthy and ubiquitous aspect of everyday life. The emergence of audible echoes is one of the factors contributing to uncomplimentary quality deterioration. Network hardware and end-user devices are intrinsically prone to this sort of quality deterioration. Designing efficient acoustic echo cancellation (AEC) devices is vital for improving listening comfort and voice quality. When we utilize inexpensive and small analog components, an echo canceller operates poorly or not at all in the system if the net nonlinear distortion is greater than a certain value. Many adaptive filters are used to remove the echo from the microphone signal to solve this problem. Nonetheless, it is difficult to accomplish the preeminent performance of the AEC in real-time circumstances. In this work, we propose nonlinear acoustic echo cancellation (NAEC) using dense long short-term memory (LSTM)-based deep learning (D^2L^2). Deep learning has been applied to the concept of speech source separation (SSS). In our deep learning based NAEC, the near-end signal is separated from the microphone using LSTM layer training. Before learning commences, the Short-Time Fourier Transform (STFT) is used to extract frequency-time domain features from the acoustic signal. In the learning part of D^2L^2, two targets are assigned. The spectral Magnitude Mask (MM) is the primary, and the Near-end Signal Mask (NSM) is the secondary mask. The simulation shows that our D^2L^2 achieves a higher Echo Return Loss Enhancement (ERLE) than other works.

1. INTRODUCTION

Hands-free communication often involves a dialogue between two speakers positioned at near-end and far-end locations [1]. The primary speech signal is captured by the near-end microphone, but it also picks up two unwanted signals: background noises and an echo caused by the loudspeaker reproducing the far-end signal. The presence of echoes, resulting from the convergence of sound waves between the speakers and the microphone output, can reduce speech intelligibility at the far-end. Numerous experiments on acoustic echo cancellation (AEC) devices, which attempt to eliminate echoes and preserve near-end speech, have been conducted to address this issue.

In recent years, nonlinear (NL) deviations in the echo pathway have been observed that are not negligible in the middle of the amplifier's emission or the far-end signal; however, these deviations are caused by the miniaturization of electrical parts used in hands-free equipment, such as wearable technologies, intelligent speakers, and smartphones. Therefore, in run-through, echo cancellation systems that assume a linear track of echo frequently fail. Different AEC algorithms for dealing with nonlinear problems have been suggested for determining nonlinear distortions on the echo route to reduce this incompatibility [2] further. Despite frequently necessitating research, the Volterra successions demonstrated victory in modeling NAEC with scrawny nonlinearities and reminiscence using nonlinear root functions. The ISF-NSFC and PRF cooperative investigation program supported this work by providing the necessary apparatus and methodological guidance with a high degree of computational complexity [3].

Hammerstein models provide a condensed form of echo cancellation. Additionally, Bayesian state-space modeling, adaptive functional link filters, and kernel-based techniques [4] are frequently applied to nonlinear AECs. Authors [5] approached the issue by taking away a frequency-to-time perspective and using sub-band adaptive filtering and multiplicative function approximation [6], effective Volterra succession modeling with annoyed band components. In contrast to conventional methods, artificial NNs offer an unconventional structure for highly nonlinear modeling [7].

For instance, nonlinear distortions on an echo track were estimated using a fully connected CNN (FCCNN) integrated with Hammerstein and adaptive filtering, as described in the previous study [8]. More recently, Halimeh et al. [9] proposed a model incorporating both Hammerstein and Wiener to predict linear and nonlinear echoes within an FCCNN. Despite yielding promising results, these methods still exhibit suboptimal performance in real-world scenarios, potentially due to two key factors. Firstly, these models inadequately capture the true nature of the distortions introduced by modern sophisticated devices onto the distant signal. Secondly, the majority of the data is parametric, requiring the pre-

determination of NL basis functions and memory durations. For instance, the models proposed in the previous study [3] assume a specific amount of remembrance knocks, and fixed nonlinear activation utilities are used inside the neural network [8, 9]. In actual installations, these flaws could lead to less-than-ideal solutions.

Adaptive filters and other signal-processing techniques are frequently used in traditional AEC approaches. However, these result in drawbacks such as higher computational complexity, instability, filter coefficients failing to converge to the optimal value, and sensitivity to signal characteristics (echo delay, SNR, reverberation). Recently LSTM networks [10] have been applied as an alternate approach for echo cancellation and suppression due to their ability to capture NL relationships between input and output signals.

This paper proposes a Dense LSTM Deep Learning D^2L^2 algorithm for nonlinear acoustic echo cancellation (NAEC) wherein the spectral magnitude plays a major role in estimating the speech source of an acoustic microphone signal. The target speech is then gradually improved by using the near-end speech detector of the binary mask as an extra source in the subsequent component for magnitude mask determination. The estimated mistakes in near-end speech can be corrected by concurrently teaching both target modules using a unique loss function MSE. Lastly, the estimated results from the spectral mask network and the near-end speech detection device of the original signal input are used to produce near-end signal. As a result, the cascading architecture makes use of each module's advantages and is expected to yield accurate magnitude and phase estimates.

D^2L^2 has three layers of BILSTM and one fully connected layer that is sandwiched between a fully connected input and output layer. The usage of multiple BILSTM layers improves the model capacity, learning more complex nonlinear mapping between input and output signal and learning of hierarchical representation of input audio signal. It also improves the learning dynamics and performance of echo cancellation.

The research contributions of this paper are as follows:

1) This research presents a novel approach to acoustic echo cancellation (AEC) using D^2L^2. In contrast to conventional approaches that depend on adaptive filters, our method models and suppresses echoes in real-time audio signals by utilizing deep learning techniques.
2) We present a unique architecture made up of several layers of Bidirectional long short-term memory (BILSTM) that is intended to capture the intricate temporal dependencies and nonlinear interactions found in echo signals. Our model learns hierarchical representations of the input audio data by stacking many BILSTM layers, which enables more precise echo cancellation.
3) To evaluate the performance of the D^2L^2-based AEC, several simulations were run. We have compared state-of-the-art adaptive filtering techniques, such as KIPNLMS, PFLAF, and NSAEC, with simulated and real-world datasets and show notable increases in echo suppression performance. Quantitative metrics ERLE and convergence time indicate superior performance across various acoustic environments and echo scenarios.
4) Our research contributes to advancing the state-of-the-art in AEC technology by harnessing the power of deep learning and LSTM networks.

The paper is structured as follows. A review of the literature is presented in Section 2. A comprehensive outline of the formulation of the proposed D^2L^2 algorithm is given in Section 3. In Section 4, the primary objective and performance evaluation of the ERLE for both the proposed method and recent works are analyzed and compared. Lastly, Section 5 deals conclusion.

2. RELATED WORK

In wireless transceiver communication models and clever speakers, the connection between the loudspeaker and the microphone produces acoustic echoes. This approach considerably lowers the effectiveness of automated speech recognition (ASR) in smart speakers and significantly worsens speech communication quality. Typical acoustic echo cancellation (AEC) techniques employ adaptive algorithms to pinpoint the loudspeaker-to-microphone impulse response (IR). In delay-sensitive circumstances, the LMS adaptive filtering algorithm [10] is frequently used.

For quick convergence and little computing burden, frequency domain LMS algorithms are frequently used [10]. Another widely used technique is the frequency-domain adaptive Kalman filter (FDKF) [11], which has recently been modified [12]. When nonlinear distortion in the acoustic echo route is not trivial, LAEC techniques perform noticeably worse [13].

To further reduce the echo, a residual echo suppression (RES) module is typically needed. Typically, RES is carried out by calculating the filter coefficients, linear AEC output, and far speech to determine the spectrum of residual echoes [14]. Nevertheless, achieving stability between near-signal distortion and remaining echo attenuation poses a challenge for Residual Echo Suppression Systems utilizing signal statistics.

The recent introduction of deep learning methods for neural networks has enabled the replication of neural network designs, including both time domain and time-frequency (TF) domain approaches. In TF domain techniques, spectral information is extracted using an STFT. A Fully Connected Network (FCN) [15] is required to accommodate multiple input signals in RES.

Another addition to AEC was the Recurrent Neural Network (RNN) constructed by either the BILSTM or the LSTM unit [16]. These techniques perform poorly because they disregard the link between the phase and magnitude, even though this method is incompetent for phase prediction in the testing environment [17]. Chen et al. [18] presented an AEC approach that entrenched the CNN architecture with full-time domain feature representation for an audio source segregation network called Conv-TasNet, and the amount produced by the adaptive filter was backed to custom many streams.

While simulations validate the benefits of incorporating supplementary signals into the network, the model's reliance on a complex network topology proves inefficient in leveraging data from multiple streams, resulting in a plethora of parameters that hinder its practical implementation. Additionally, further research on commercial loudspeakers is necessary to substantiate the advantages of multistreams. Table 1 outlines the latest techniques in acoustic echo cancellation.

Table 1. Literature works of NAEC

Year	Method	Stability	Convergence	Computational Complexity	Remarks
2004	Interpolated frequency domain sampling rate conversion (SRC)	Low Steady	Sluggish	Low	Voice chat audio playback has an impact on AEC. Corrected only the playback sampling rate; the sample rate offset from the microphone is not taken into account.
2008	Predistorter	Steady	Rapid	Low	Predistorter included without NL port, the system acts exactly like a linear system. Thus, whichever AEC method is utilized to eradicate the echo.
2010	Filter Model with Shortening	Ascetically Steady	Rapid	Hinge on the Filter method	A shortening filter drastically reduces the computational complexity by reducing the room impulse response.
2011	Hammerstein NL design	Steady	Sluggish	high	The linear component of the LEM system is modeled using a linear section, while the NL is modeled utilizing the NL section.
2011	DCAF (Drift-Compensated Adaptive Filtering)	Ascetically Steady	Rapid	Moderate	To enhance speech when the target speech is distorted by asynchronous interference.
2012	An Improved Proportionate NLMS Algorithm Based on The 10 Norm	Steady	Rapid	High	Utilize the sparsity of the system that must be recognized by using the l0 norm. It has a higher rate of convergence than the NLMS method, is still practical and reliable to use, and does not have issues with fixed-point instability.
2013	Nonlinear Cascaded Filter	Ascetically Steady	Rapid	Low	This approach separates the nonlinear from the linear components, ensuring constant stability and quick convergence.
2014	Modified version of Partitioned Block Frequency Domain Adaptive Filter (PBFDAF)	Steady	Rapid	High	AEC Complexity and Delay does not consider the issue of sampling rate incongruity with filter inputs.
2014	Proportionate functional link adaptive filter	Steady	Rapid	High	PFLAF is a proportional adaptation variant of SFLAF, which is based on a sparse representation of functional links that updates the coefficients for nonlinear modeling. The performance of PFLAF is superior in comparison with NLMS and SFLAF.
2017	Volterra Filters	Steady	Sluggish	Depends on the Adaptive filters used	Adaptive algorithms of any kind can be used to update the coefficients. The adaptive algorithm of choice determines how difficult the process is.
2020	Harmonic Distortion echo suppressor	Steady	Rapid	Low	Because of its quicker convergence and lesser computing complexity, this may primarily be employed for handheld devices.
2020	Kernelized improved proportionate NLMS algorithm	Steady	Rapid	High	It is the kernelized variant of IPNLMS. The nonlinear echo path is easily modeled, and the global minimum is easily attained using the kernel method. KIPNLMS algorithm performs better than KLMS, KNLMS and KPNLMS.
2021	PercepNet joint noise and residual echo suppressor	Steady	Rapid	Moderate	Integrating a conventional acoustic echo canceller with a deep neural network (DNN)/hybrid signal processing-based low-complexity combined residual echo and noise suppressor.
2022	Neural Cascade Architecture	Highly Steady	Rapid	Depends on the Sequence Length of the signals	The suggested cascade structural design is accomplished from top to bottom with only one loss utility rather than utilizing sequential training phases with discrete loss functions.
2022	Nonlinear stereophonic acoustic echo cancellation	Steady	Rapid	High	NSAEC follows a functional link-based adaptive filtering approach and uses a sub-filter approach to enhance the convergence. NSAEC has superior performance compared to KIPNLMS and PFLAF.

3. PROPOSED D^2L^2-DENSE LSTM DEEP LEARNING METHODOLOGY

3.1 System model

The nonlinear AEC's system model is shown in Figure 1. The room impulse response (RIR) of h(n) is combined with the far-end signal, x(n), to produce an echo signal that is mixed with the near-end signal of s(n). The three components of echo, near-end, and noise signals construct the microphone signal y(n), as shown by:

$$y(n) = s(n) + d(n) + v(n) \quad (1)$$

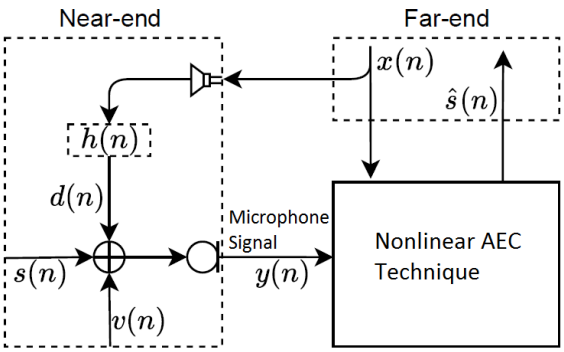


Figure 1. Nonlinear acoustic echo cancellation system

The AEC's goal is to reduce echo d(n) and transfer near-end signal s(n) towards the far-end. Adaptive filters are utilized in classical AEC methods to estimate the IR $\hat{h}(n)$, where y(n) represents the microphone signal and x(n) represents the far-end signal. The estimated signal $\hat{d}(n)$ of the echo is then calculated as follows:

$$\hat{d}(n) = \left(\hat{h}(n)\right)^T \mathrm{x}(n) \tag{2}$$

where, x(n) is the input buffered, $\hat{h}(n)$ is the adaptive filter of length L, and T denotes transpose. The microphone signal is subtracted from $\hat{d}(n)$ to yield the system output. This output signal or the error signal e(n) is defined as:

$$e(n) = y(n) - \hat{d}(n) \tag{3}$$

The estimation of the acoustic route h(n) is the crucial step in AEC methods. Traditional AEC techniques cannot manage nonlinear aberrations because they are inherently linear systems. Fast transversal filters [19], adaptive Volterra filters [20], Subband filters [21], and Functional link AF [22] explicitly represent the AEC system's nonlinearity. However, in actual situations with large nonlinearities and potent interference signals, their performance is constrained.

The AEC's ultimate objective is to eliminate both the end signal of the far-end and the circumstantial noise at the end of the near-end, allowing only the signal at the near-end to be sent to the far-end. As per speech separation [23], AEC can be conceptualized as a separation issue, with the primary objective being the isolation of the near-end speech, which constitutes the main source of interference in the microphone recording. To address this, we utilize controlled speech separation to process the microphone inputs and simultaneously extract both the original near-end signal and far-end signal, instead of directly predicting the acoustic echo path. The clear near-end signal from the microphone signal is extracted using a dense LSTM architecture that is based on deep learning and is covered in the next section.

3.2 D^2L^2 dense LSTM deep learning-based NAEC

Here we incorporate a deep learning approach into NAEC for audio echo cancellation. Figure 2 is a diagram that illustrates our suggested method. The schematics use a fully connected input layer with 1024 units that passes the signal to three BILSTM layers with 512 units in each layer and a fully connected layer with 600 units. The signal is then passed on to an output layer with 512 units.

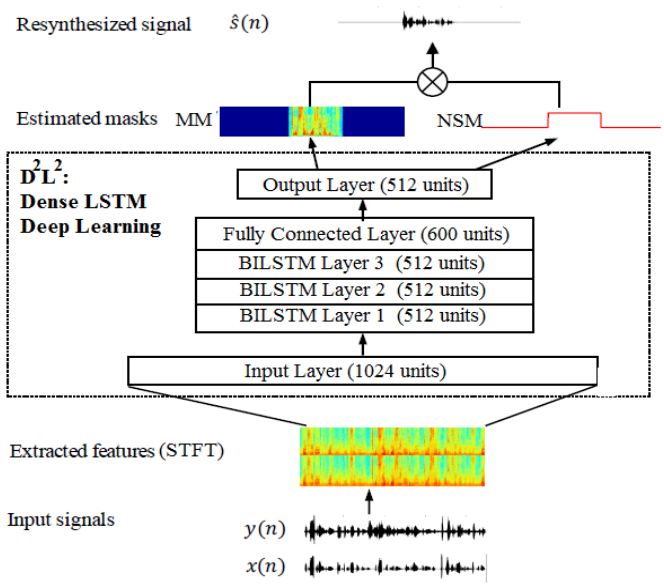


Figure 2. Proposed D^2L^2: Dense LSTM deep learning based NAEC

Broadly, the approach aims to mitigate background noise and acoustic echo, facilitating the extraction of near-end speech. This involves estimating both the magnitude mask (MM) for the microphone spectral signal and the binary mask for near-end speech (NSM) from $x(n)$ and $y(n)$. More precisely, a dense LSTM receives acoustic features derived initially from both the microphone signal and the far-end signal. To estimate the MM and NSM, the top layer of the LSTM functions as a mask estimation layer. The near-end signal estimate is then resynthesized using the mask estimate to the microphone signal.

3.3 STFT-based spectral feature extraction

Generally, speech signals are not stationary. Often, we want to examine the spectrum of each phoneme independently, but if we convert a spoken phrase to the frequency domain, we acquire a spectrum that is an average of all phonemes in the sentence. We can concentrate on the signal's characteristics at a certain moment by segmenting the signal into shorter chunks. We acquire the STFT of the signal by windowing and taking the Discrete Fourier transform (DFT) of each window.

By selecting an appropriate window function, a time-domain STFT partitions an acoustic wave into segments, revealing how frequency components evolve. An advantage of STFTs is that their parameters possess a physical and intuitive interpretation, corresponding directly to the spectrum.

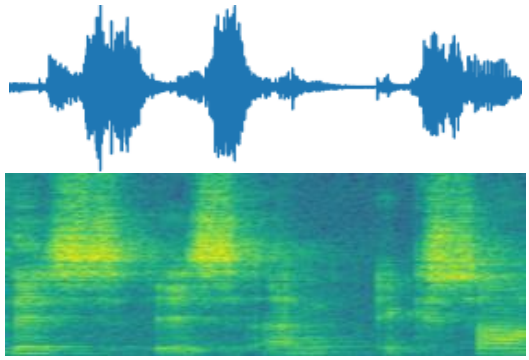


Figure 3. STFT spectrogram of the signal

Specifically, for an input signal s(n) and window w(n), the transform is defined as,

$$S(m,k) = \sum_{n} s(m,n)\, e^{-\frac{j2\pi nk}{N}} \tag{4}$$

$$s(m,n) = s(n) * w(n - mL) \tag{5}$$

where, the fast Fourier transform (FFT) size is represented in terms of 2 orders and denoted here as N, n is the sample time index within the given frame, m is the frame index and w(n) is the windowing callback with N signals. w(n) is positioned at shifting step size L, and $1 - \frac{L}{N}$ is the overlay ratio among continuous frames. The input signals are presented into 16 ms frames with an 8 ms delay between successive frames after being sampled at 16kHz. Every frame is then subjected to a 512-point STFT to obtain a vector of spectral magnitudes, which results in 256 frequencies. To form the input feature vector, we concatenate the STFT features extracted from the far-end signal $F_x(t)$ and the microphone signal $F_y(t)$ as follows:

$$F(t) = \left[F_y(t), F_x(t)\right]^T \tag{6}$$

where, the feature vectors of the microphone signal and far-end speech at time frame m are indicated by the symbols $F_y(t)$, $F_x(t)$, as a result, 1024 (512×2) is the input's dimensionality. The STFT spectrogram is shown in Figure 3.

3.4 D²L² network design

RNN is a type of model for recording chronological data and is used to tackle a variety of issues, including speech recognition [24], machine translation [25], and natural language processing [26]. The sequential data have long-term dependence, which leads to disappearing and expanding gradient difficulties. This problem can be resolved by the application of LSTM and gated RNNs. In the LSTM network, supplementary memory cells and gate procedures exist. At the time of backpropagation, moving out of the gradient will be precluded by the gate operations. With fewer gate functions, the gated recurrent unit (GRU) performs similarly [27].

Gated commentary: Two successive time steps are fully connected by an RNN. The depths of the gated feedback (FB), feedforward (FF), and recurrent feedback may all be modeled together since their connections to the model with orthogonal depths do not overlap. We suggest an attention gate with these three features, which regulates the flows from each state to improve the overall performance via D²L². Figure 4 depicts deep learning methodology and structure.

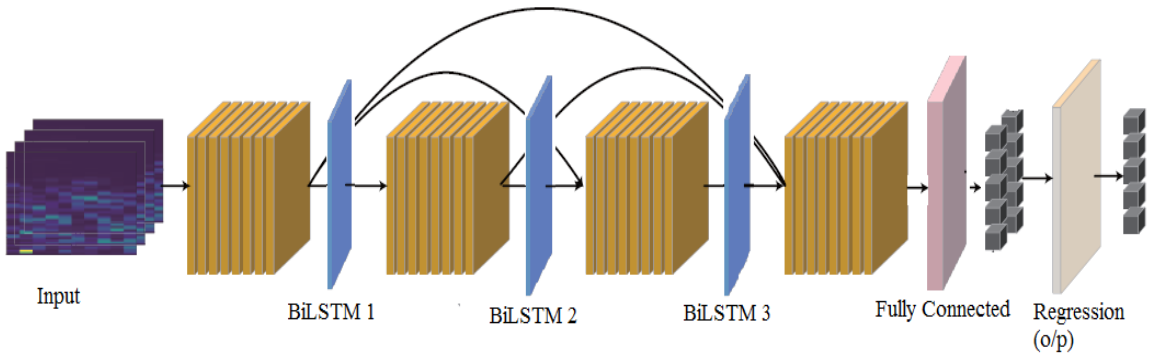


Figure 4. Proposed LSTM

3.5 LSTM-based deep learning

In the general structure of the recurrent layer, the preceding hidden state h_{t-1} and input x_t are used to define each hidden state at time t as h_t, which is given by:

$$h_t = \phi(h_{t-1}, x_t) = \phi(W x_t + U h_{t-1}) \tag{7}$$

where, ϕ is the "tanh" nonlinear process with each sample operation, W is the recurrent layer weight matrix and U is the feed-forward weight matrix. For a typical RNN architecture, the last hidden state h_{t-1} will store the previous state inputs for further calculations. This leads to poor memory processing when the hidden state is allocated less storage space, which is reflected in the continuity of the state variables. To overcome this problem, a stacked RNN is employed to capture prolonged dependencies across multiple state statuses within the hidden layers, as illustrated below:

$$h_t^j = \phi\left(W^j h_t^{j-1} + U^{j\to j} h_{t-1}^j\right) \tag{8}$$

where, U^j is the changeover from time step t-1 to time step t at layer j, which impacts the weight matrix, and W^j is the transition from layer j-1 to layer j, which affects the weight matrix. The sequential data may be modeled across various timeframes using the stacked RNN. When prediction data are stacked at the top layer, long-term (LT) dependencies are highly covered by hidden state storage. Hence, deep recurrent network formation is necessary by connecting the current hidden state to the previous ones. Thus, we can improve the long-term dependency [28], and it is represented as:

$$h_t^j = \phi(W^j h_t^{j-1} + \sum_{k=1}^{K} U^{(k,j)\to j} h_{t-k}^j) \tag{9}$$

where, K is the recurrent depth and $U^{(k,j)\to j}$ is the weight matrix of layer j with the transition occurring from time t-k to time t. The shortcut routes from many earlier concealed states are made by direct linkages. The model with shortcut pathways allows access to earlier hidden states farther away from h_t^j with the same number of transitions as the model without shortcut paths. Many recurrent models typically establish connections solely among hidden states within the

same layer. However, the network layer adaptly mitigates numerous timing issues by aggregating feedback connections from preceding hidden states h_{t-1}^{j} to the hidden state h_t^j across different layers of h_t^j, as demonstrated below:

$$h_t^j = \phi\left(W^j h_t^{j-1} + \sum_{i=1}^{L} U^{i\to j}\, h_{t-1}^{i}\right) \tag{10}$$

where, the feedforward depth is L and $U^{i\to j}$ is the weight matrix transition from time t-1 to t. The global gate is designed to regulate the number of streams between diverse hidden states with diverse time measures [27].

$$h_t^j = \phi\left(W^j h_t^{j-1} + \sum_{i=1}^{L} g^{i\to j}\, U^{i\to j} h_{t-1}^{i}\right) \tag{11}$$

The gate controlling the flow of information from each hidden state (HS) to multiple preceding HSs across all layers is denoted as $g^{i\to j}$.

$$g^{i\to j} = \sigma\left(w_g h_t^{j-1} + u_g^{i\to j} h_{t-1}^{*}\right) \tag{12}$$

where, h_t^{j-1} is the same size as the state vector w_g and h_{t-1}^{*} is the size of the state vector $u_g^{i\to j}$. It has been established that for a concatenated vector comprising all hidden states from the mentioned time step, the sigmoid function applied elementwise to the vector is denoted by *.

In gated LSTM, the output gate o_t^j, forget gate f_t^j, input gate i_t^j, and memory cell gate $\tilde{c}_t^j$ are defined as follows:

$$i_t^j = \sigma\left(W_i^j h_t^{j-1} + U_i^{j\to j} h_{t-1}^{j}\right) \tag{13}$$

$$f_t^j = \sigma\left(W_f^j h_t^{j-1} + U_f^{j\to j} h_{t-1}^{j}\right) \tag{14}$$

$$o_t^j = \sigma\left(W_o^j h_t^{j-1} + U_o^{j\to j} h_{t-1}^{j}\right) \tag{15}$$

$$\tilde{c}_t^j = \phi\left(W_c^j h_t^{j-1} + \sum_{i=1}^{L} g^{i\to j}\, U_c^{i\to j} h_{t-1}^{j}\right) \tag{16}$$

Like in conservative LSTM, the memory cell gate $\tilde{c}_t^j$ has a feedback loop for gates in gated LSTM. Using gates and the memory cell state in (13)-(16), the new hidden state h_t^j and new memory cell state c_t^j are modeled as

$$c_t^j = f_t^j . c_{t-1}^j + i_t^j . \tilde{c}_t^j \tag{17}$$

$$h_t^j = o_t^j . \phi\left(c_t^j\right) \tag{18}$$

where, the default dot product denotes matrix multiplication.

3.6 Modelling of D²L²

Deep learning networks can be trained more effectively when skip-linking state connections are incorporated as they enable connections to bypass certain layers. This approach which involves Skip linking state connections to feedback connections has been utilized in various studies [29, 30]. We apply skip-linking state connections to recurrent connections in this work. The skip-linking state connections over time are equivalent to the shortcuts from the various HSs that came before to the hidden state at time t. The feedback linkages between several layers of various time steps are among the shortcut routes. The attention gate normalizes each connection, resembling closely the global gated feedback RNN. The dense recurrent neural network is represented as follows:

$$h_t^j = \phi(W^j h_t^{j-1} + \sum_{k=1}^{K}\sum_{i=1}^{L} g^{(k,i)\to j}\, U^{(k,i)\to j} h_{t-k}^{i}) \tag{19}$$

where, the attention gate is mathematically denoted as $g^{(k,i)\to j}$ below:

$$g^{(k,i)\to j} = \sigma(w_g^i h_t^{j-1} + u_g^{(k,i)\to j} h_{t-k}^{i}) \tag{20}$$

where, (20) is a gate callback of the foregoing HS at time t-k and layer i, and (12) is a callback of all the interconnected foregoing HSs. The dense recurrent neural network is straightforwardly extended to dense LSTM and is described as follows:

$$i_t^j = \sigma(W_i^j h_t^{j-1} + \sum_{k=1}^{K}\sum_{i=1}^{L} g_i^{(k,i)\to j}\, U_i^{(k,i)\to j} h_{t-k}^{i}) \tag{21}$$

$$f_t^j = \sigma(W_f^j h_t^{j-1} + \sum_{k=1}^{K}\sum_{i=1}^{L} g_f^{(k,i)\to j}\, U_f^{(k,i)\to j} h_{t-k}^{i}) \tag{22}$$

$$o_t^j = \sigma(W_o^j h_t^{j-1} + \sum_{k=1}^{K}\sum_{i=1}^{L} g_o^{(k,i)\to j}\, U_o^{(k,i)\to j} h_{t-1}^{i}) \tag{23}$$

$$\tilde{c}_t^j = \phi(W_c^j h_t^{j-1} + \sum_{k=1}^{K}\sum_{i=1}^{L} g_c^{(k,i)\to j}\, U_c^{(k,i)\to j} h_{t-1}^{i}) \tag{24}$$

In contrast to the gated FB-LSTM, the parameter g is distributed across all memory cell states and gates. We examined the spontaneous adjustment of dense connections. Figure 5 illustrates the state connection diagram of dense LSTM. Figure 5 (a) shows the conventional LSTM unfolded in time, and Figure 5 (b) represents the gated feedback LSTM. Figure 5 (c) shows the preceding state connections across the LSTM, and Figure 5 (d) shows the dense LSTM related with Figure 5 (b) and Figure 5 (c). HSs are marked in red. The links of dense LSTM utilized in FF steps are highlighted in bold. The FB-states between the higher and inferior layers are represented in yellow.

3.7 D²L² training targets

For training the proposed novel D²L² algorithm, training inputs and training targets are constructed. The extracted STFT magnitude features are the input for training in D²L². There are mainly two training targets. One of the targets is the MM of spectral extracted far-end and microphone signals. The second target is the NSM.

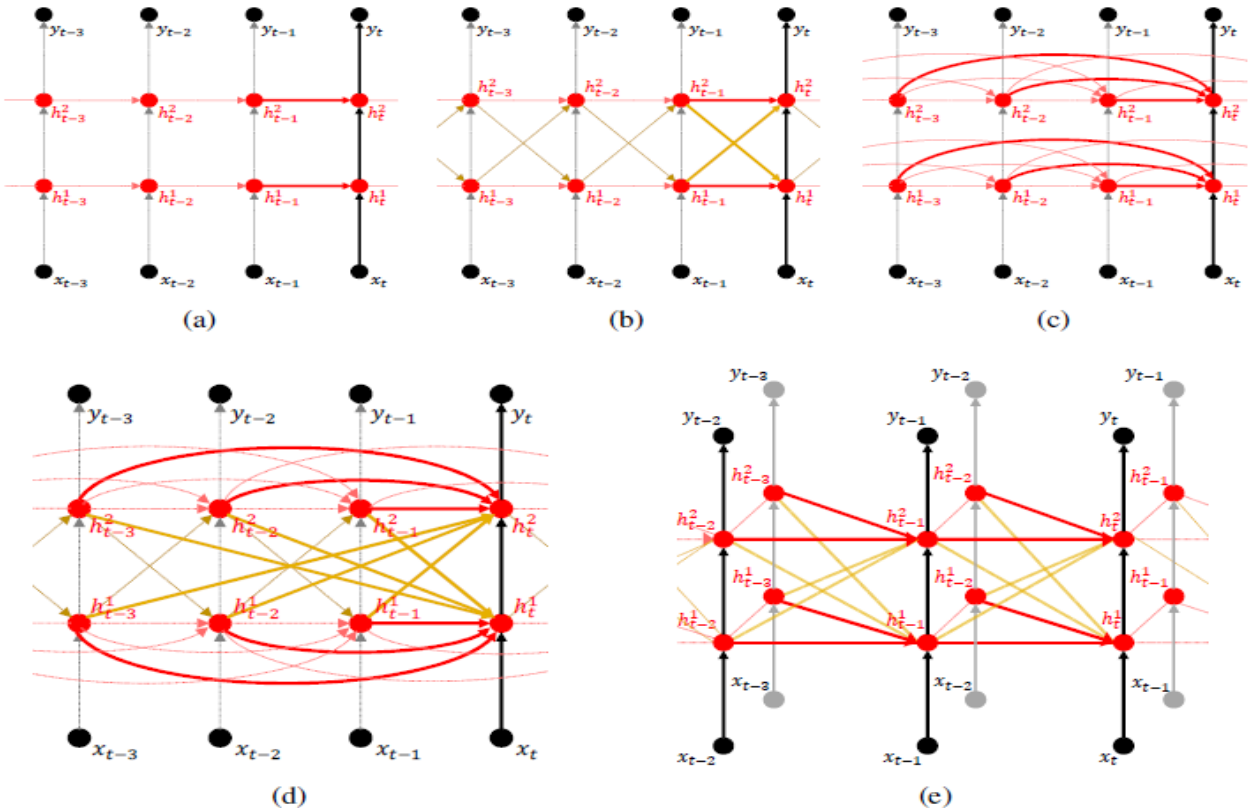


Figure 5. Dense LSTM state representations

3.7.1 Magnitude mask (MM)

The MM, which is defined as follows, is utilized as the training objective for near-end speech extraction.

$$MM = \sqrt{\frac{S^2}{Y^2}} \tag{25}$$

S^2 is the spectral magnitude of the clean signal, and Y^2 is the spectral magnitude of the microphone signal.

3.7.2 Near-end signal mask (NSM)

To improve the predicted MM better, NSM is used. NSM is a binary detector based on streams that show whether nearby signals are present. In contrast to the Dual Talk Detector (DTD), which assesses simultaneous interaction between near-end and far-end speakers, the proposed NSM focuses solely on near-end speech.

$$\text{NSM}(t) = \begin{cases} 1, if \max_f |S(t,f)| > 0 \\ 0\,, \qquad\qquad else \end{cases} \tag{26}$$

For each sample stream, the microphone signal's echo and background noise estimation is further refined, excluding near-end speech while ensuring the predictability of near speech by the MM in other frames. Subsequently, the estimated NSM is employed on the estimated MM, serving as a residual echo suppressor with supervision.

The ultimate propulsion mask of the dense LSTM is M=MM*NSM, which can also be expressed in polar coordinates:

$$\begin{cases} M_{mag} = M \\ M_{phase} = arctan2(M) \end{cases} \tag{27}$$

The projected near speech $\hat{S}$ will be calculated by,

$$\hat{S} = Y_{mag} \,.\, M_{mag} \,.\, e^{Y_{phase} + M_{phase}} \tag{28}$$

The value range of the training objectives ranges from zero to one. The activation cascade is a sigmoid function at the output layer of the regression. BiLSTM is trained by the loss function MSE which is the mean square error between S and $\hat{S}$ and solved with the Adam optimizer [31].

The inverse STFT receives the phase of the microphone signal and the estimated spectral magnitude of the near-end speech to produce a near-end waveform signal estimation.

4. SIMULATION SETUP

For this deep learning-based NAEC evaluation, the AEC Challenge 2023 [32] dataset is utilized. These datasets include a synthetic dataset as well as synthetic recordings from over 10,000 genuine audio equipment and real-world settings containing human speakers.

Four different signal types namely near-end speech, background noise, far-end speech, and matching echo signals must be provided to train the network. The official synthetic dataset for near-end speech s(n) has 10,000 utterances; we chose the first 500 utterances as the test set, which is excluded from the training. A total of 9,500 additional utterances were used for practice. Table 2 provides a comprehensive overview of the dataset. For these multiple scenarios, we use the MATLAB simulation tool version 2020a. Table 3 provides the configuration of the simulation environment.

For each signal, 16000 samples were tested, and four signal combinations were taken for the process: far-end signal, near-end signal, echo signal, and microphone signal. The inputs

given to dense LSTM are the microphone signal and far-end spectral magnitude.

Table 2. Simulation environment

Parameters	Value
Dataset	
Testing Samples	500
Training Samples	9500
Each Sample Archive	Near-end signal, Far-end signal, Microphone signal, Echo signal
STFT	
Sampling Rate	16KHz
Window Length	16ms
Overlap	8ms
Size	512
Noise-Symbol Energy	1 MHz
D^2L^2: Dense LSTM Deep Learning	
Number of Epochs	50
Minimum Batch Size	32
Solver Optimizer	ADAM
Learning Rate	0.001
Learning Rate Schedule	Piece Wise
Learning Rate Drop Factor	0.9
Learning Rate Drop Period	1 s
L2 Regularization Factor	0.0001
Activation Function	Stochastic Gradient Descent
Momentum	0.9
Gradient Decay Factor	0.9
Dominator Offset	10^{-8}
Execution Environment	CPU

Table 3. Configuration of simulation environment

AEC-Challenge 2023		
Key Words	**Real Datasets**	**Synthetic Datasets**
Number of Recordings	50000	120000
Number of Different Speakers	**10000c**	**1627**
Recording Platforms	Microsoft Windows, Android Devices	LibriVox Project [9], Free sound and DEMAND
Audio Mode	WASAPI raw audio	ITU-T P.808
Sampling Rate	48 KHz	16 KHz
Audio Duration	10 sec	10 sec
Sentence List Source	TIMIT [15]	LibriVox Project [9]
Noise Range	Real-time	0-40 dB
Echo Ratio	Real-time	-10dB to 10dB
RT60	Desktop- 4387 Mobile-1251	200ms to 1200ms
Gender Clips	Male- 500, Female-500	Randomly selected
Number of Scenarios	6	5
	1. Without path change of echo, single talk far-end	1. Single Talk
	2. With path change of echo, single talk far-end	2. Double Talk
	3. Without path change of echo, single talk near-end	3. Near-end Noise
	4. Without path change of echo, double talk	4. Far-end Noise
	5. With path change of echo, double talk	5. Nonlinear Distortions
	6. RT60 estimation by Sweep signal	

5. RESULTS AND DISCUSSION

In this section, we experiment with our proposed D^2L^2 algorithm for NAEC and compare it with previous adaptive filter-based echo cancellation methods from recent works, such as NSAEC, KIPNLMS, and PFLAF [33-35]. The key metrics for AEC's performance are the convergence time and the echo return loss enhancement (ERLE).

ERLE variation is frequently used to assess the performance of AEC systems when there is background noise, and it is defined as:

$$ERLE = 10\ \log_{10}[\frac{\sum_n y^2(n)}{\sum_n \hat{S}^2(n)}] \quad (29)$$

It measures the reduction in echo after applying the cancellation algorithm.

The cross-validation technique is used to assess the stability and generalization performance of the algorithm. This is done by dividing the data set into training and validation sets multiple times. The training set is used to train the model, while the validation set is used to evaluate its performance.

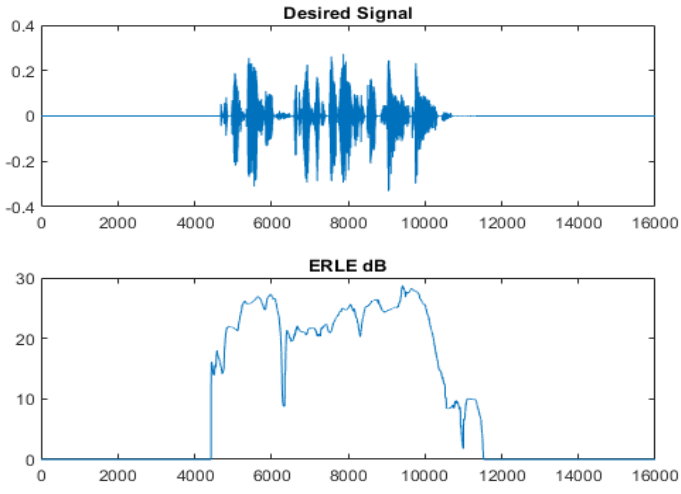
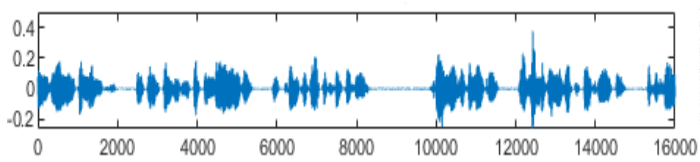
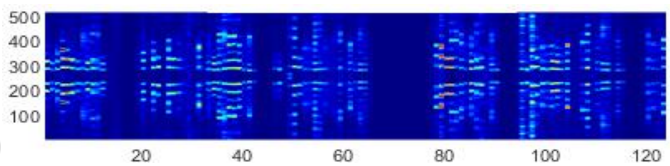
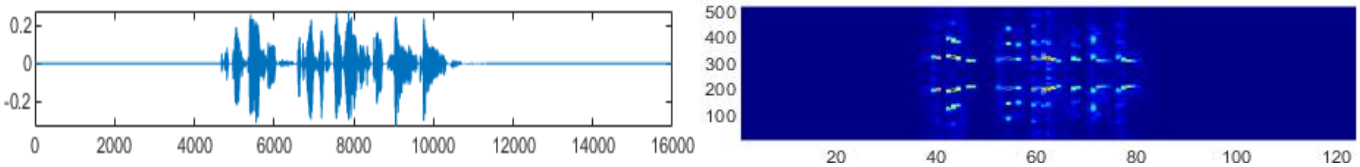


Figure 6. Estimation of ERLE for an echo canceled signal

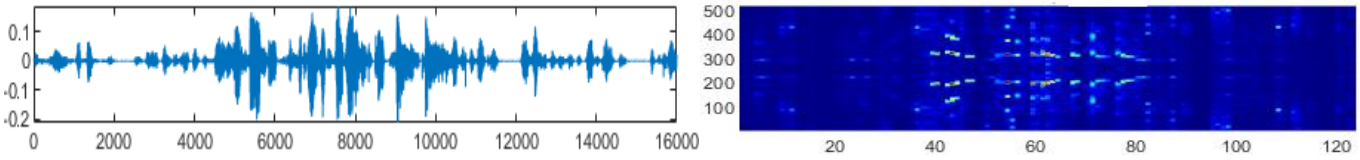




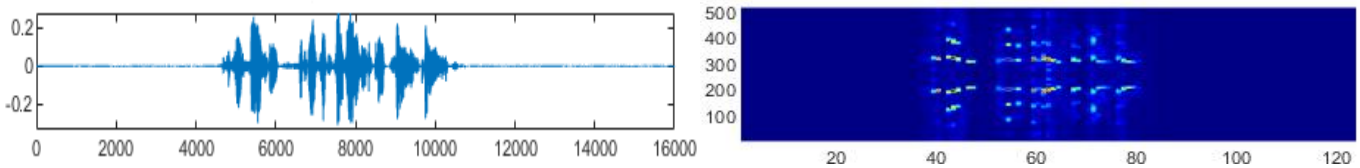
a) Far-end Signal and spectrogram



b) Near-end Signal and spectrogram



c) Microphone Signal and spectrogram



d) Echo Cancelled Signal and spectrogram

Figure 7. Non-linear echo cancellation results of proposed D^2L^2

Figure 6 depicts the ERLE performance in the presence of background noise for the corresponding desired signal of the near-end. This finding led to the development of the D^2L^2 algorithm to abate the influence of nonlinear distortion despite predicting the echo cancellation signal spectrum. The inverse STFT causes the magnitude of the spectrum signal to increase to the real signal of the near-end to recover.

Figure 7 depicts the signal and spectrograms of test samples of far-end, near-end, and microphone signal in circumstances with nonlinear distortions, background noise, and double-talk in Figure 7 (a), Figure 7 (b), and Figure 7 (c) respectively.

The predicted outcomes of the echo cancellation signal under consideration are shown in Figure 7 (d). The suggested technique overcomes the best echo clampdown and leaves the least amount of noise and echo in the recovered signal.

Figure 8 demonstrates the improvement in the ERLE achieved by the proposed D^2L^2 algorithm compared with that achieved by the NSAEC, KIPNLMS, and PFLAF methods for signals without background noise. A maximum of 29dB for dense LSTM and 21, 16.5, and 15dB of ERLE for the NSAEC, KIPNLMS, and PFLAF methods, respectively, were obtained from implementation. In Figure 9, noise is applied, and the evaluations are verified by the ERLE for all the methods. As the SNR increases, the ERLE improves and reaches 36.5 dB for Dense LSTM and 19 dB for the PFLAF method at the peak location of the signal.

Figure 10 shows the ERLE when the SNR is set to 20 dB. With the dense LSTM touching 40.3 dB, the NSAEC attained 29.7 dB of ERLE, which is more than the 10 dB improvement in D^2L^2.

In Figure 11 and Figure 12, we exemplify the ERLEs at SNR=30 dB and SNR=40dB in that order. With an increase in the SNR of 10dB, the ERLE increases to 8dB, 5dB, 4dB, and 3dB for DenseLSTM, NSAEC, KIPNLMS, and PFLAF, respectively.

In Table 4 (a) and Table 4 (b), we show the numerical comparison of the ERLE and computational time complexity for distinct SNRs in the range of 0-40dB. According to the base concept, when SNR intensification occurs, the ERLE escalates. Similarly, as shown in Table 4 (a), the proposed dense LSTM network reaches the highest ERLE of 51.8dB at an SNR of 40dB. This value is nearly 15dB greater than that of the NSAEC, 22dB greater than that of the KIPNLMS, and 26dB greater than that of the PFLAF. For PALAF, there is a 100% improvement in the ERLE for the proposed D^2L^2 algorithm.

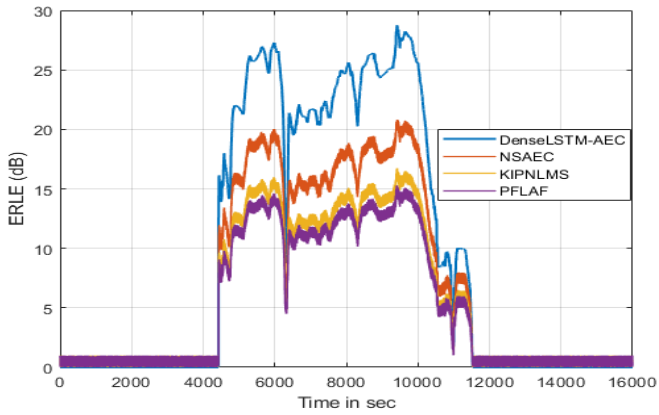


Figure 8. ERLE measure when no background noise

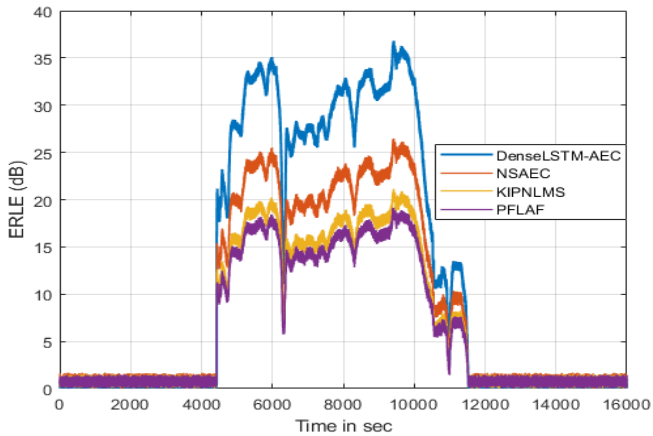


Figure 9. ERLE measure with SNR 10dB

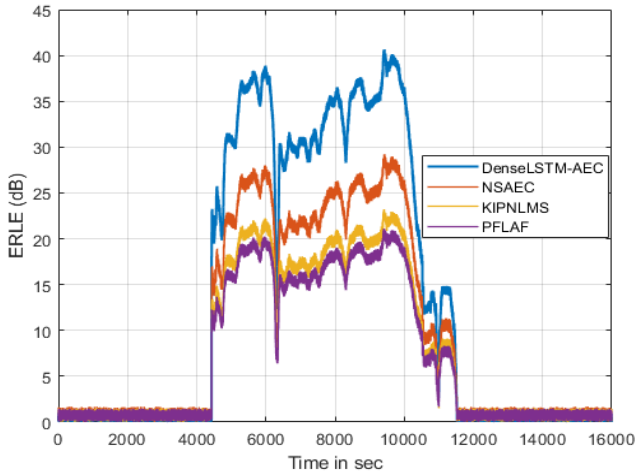


Figure 10. ERLE measure with SNR 20dB

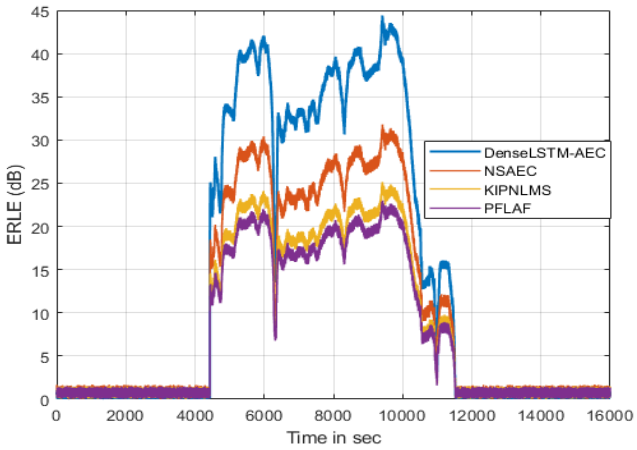


Figure 11. ERLE measure with SNR 30dB

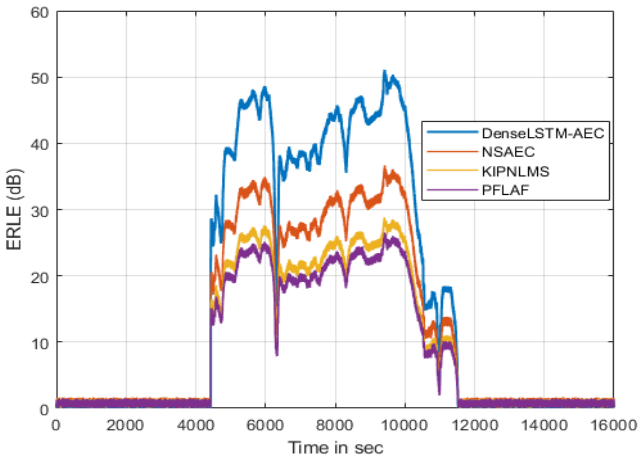


Figure 12. ERLE measure with SNR 40dB

Table 4. Comparison of ERLE

(a) Assessment of Peak ERLE(dB) with different SNR

SNR (dB)	D^2L^2: Dense LSTM	NSAEC	KIPNLMS	PFLAF
0	29	21	16.5	15
10	36.5	26	20.6	19
20	40.3	29.7	24.1	21
30	44.6	31.3	25	23.2
40	51.8	36	28.4	26

(b) Assessment of computation time(s) with different SNR

SNR (dB)	D^2L^2: Dense LSTM	NSAEC	KIPNLMS	PFLAF
0	25.39	33.15	38.78	35.12
10	26.01	32.69	39.11	37.34
20	25.84	34.88	39.35	36.09
30	24.93	32.73	38.90	38.02
40	27.38	34.04	37.88	35.59

The time complexity is analyzed to prove the superiority of the proposed algorithm with respect to the overall factors. The computational times of the methods are depicted in Table 4 (b). On average, the D^2L^2 algorithm needed 25 sec of computation time, which was 7 sec, 11 sec, and 12 sec less than that needed by the NSAEC, KIPNLMS, and PFLAF methods for echo cancellation, respectively.

Table 5. Comparison of relative mean ERLE

SNR (dB)	LMS	NLMS	KIPLMS	PFLAF	NSAEC	D^2L^2
0	0.59	0.61	1	0.65	1.57	2.12
10	0.53	0.57	1	0.61	1.44	1.94
20	0.58	0.66	1	0.75	1.44	1.94
30	0.49	0.51	1	0.88	1.53	2.06
40	0.52	0.54	1	0.91	1.51	2.04

In Table 5 mean ERLE of various algorithms is compared relatively against KIPLMS. From Tables 5 and 4 (b) it can be seen that the D^2L^2 approach has considerable improvement in echo cancellation performance with computational time less than other algorithms.

6. CONCLUSIONS

This work revealed that our proposed project D^2L^2: DenseLSTM-based deep learning method for nonlinear acoustic echo cancellation has low time complexity and high accuracy even for configured noises and distortions. The results are highlighted by comparing the performance of our method with those of earlier methods. We substantiated that the phase and magnitude spectral data are highly meritorious when applied through the magnitude mask and NSM prediction from the D^2L^2 module. Investigational outcomes in NLD settings and circumstantial noise conditions were also tested, revealing that our algorithm is effective in acoustic echo environments. In future work, we will focus on implementing time domain signal speech source separation without the STFT method.

REFERENCES

[1] Mossi, M.I., Evans, N.W., Beaugeant, C. (2010). An assessment of linear adaptive filter performance with nonlinear distortions. In IEEE International Conference on Acoustics, Speech, and Signal Processing, Dallas, TX, USA, pp. 313-316. https://doi.org/10.1109/ICASSP.2010.5495904

[2] Scarpiniti, M., Comminiello, D., Parisi, R., Uncini, A. (2011). Comparison of Hammerstein and Wiener systems for nonlinear acoustic echo cancelers in reverberant environments. In 2011 17th International Conference on Digital Signal Processing (DSP), Corfu,

Greece, pp. 1-6. https://doi.org/10.1109/ICDSP.2011.6004959.
[3] Comminiello, D., Scarpiniti, M., Azpicueta-Ruiz, L.A., Arenas-Garcia, J., Uncini, A. (2013). Functional link adaptive filters for nonlinear acoustic echo cancellation. IEEE Transactions on Audio, Speech, and Language Processing, 21(7): 1502-1512. https://doi.org/10.1109/TASL.2013.2255276
[4] Diana, D.C., Carline, M.J. (2023). Hybrid metaheuristic method of ABC kernel filtering for nonlinear acoustic echo cancellation. Applied Acoustics, 210: 109443. https://doi.org/10.1016/j.apacoust.2023.109443
[5] Van Vaerenbergh, S., Azpicueta-Ruiz, L.A., Comminiello, D. (2016). A split kernel adaptive filtering architecture for nonlinear acoustic echo cancellation. In 2016 24th European Signal Processing Conference (EUSIPCO), Budapest, Hungary, pp. 1768-1772. https://doi.org/10.1109/EUSIPCO.2016.7760552
[6] Zhang, S., Zheng, W.X. (2017). Recursive adaptive sparse exponential functional link neural network for nonlinear AEC in impulsive noise environment. IEEE Transactions on Neural Networks and Learning Systems, 29(9): 4314-4323. https://doi.org/10.1109/TNNLS.2017.2761259
[7] Zhang, H., Tan, K., Wang, D. (2019). Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions. Interspeech, 4255-4259. https://doi.org/10.21437/Interspeech.2019-2651
[8] Shi, K., Ma, X., Zhou, G.T. (2008). Acoustic echo cancellation using a pseudo coherence function in the presence of memoryless nonlinearity. IEEE Transactions on Circuits and Systems I: Regular Papers, 55(9): 2639-2649. https://doi.org/10.1109/TCSI.2008.920114
[9] Halimeh, M.M., Huemmer, C., Kellermann, W. (2019). A neural network-based nonlinear acoustic echo canceller. IEEE Signal Processing Letters, 26(12): 1827-1831. https://doi.org/10.1109/LSP.2019.2951311
[10] Zhang, H. (2022). Deep learning for acoustic echo cancellation and active noise control [Doctoral dissertation, Ohio State University]. OhioLINK Electronic Theses and Dissertations Center. http://rave.ohiolink.edu/etdc/view?acc_num=osu1650477901420567.
[11] Fan, W., Chen, K., Lu, J., Tao, J. (2019). Effective improvement of under-modeling frequency-domain Kalman filter. IEEE Signal Processing Letters, 26(2): 342-346. https://doi.org/10.1109/LSP.2019.2890965
[12] Desiraju, N.K., Doclo, S., Buck, M., Wolff, T. (2019). Online estimation of reverberation parameters for late residual echo suppression. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28: 77-91. https://doi.org/10.1109/TASLP.2019.2948765
[13] Valero, M.L., Mabande, E., Habets, E.A. (2014). Signal-based late residual echo spectral variance estimation. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, pp. 5914-5918. https://doi.org/10.1109/ICASSP.2014.6854738
[14] Carbajal, G., Serizel, R., Vincent, E., Humbert, E. (2018). Multiple-input neural network-based residual echo suppression. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, pp. 231-235. https://doi.org/10.1109/ICASSP.2018.8461476
[15] Zhang, H., Wang, D. (2018). Deep learning for acoustic echo cancellation in noisy and double-talk scenarios. Training, 161(2): 322. http://dx.doi.org/10.21437/Interspeech.2018-1484
[16] Zhang, C., Zhang, X. (2020). A robust and cascaded acoustic echo cancellation based on deep learning. In INTERSPEECH, pp. 3940-3944. https://doi.org/10.21437/Interspeech.2020-1260
[17] Luo, Y., Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(8): 1256-1266. https://doi.org/10.1109/TASLP.2019.2915167
[18] Chen, H., Chen, G., Chen, K., Lu, J. (2021). Nonlinear residual echo suppression based on dual-stream DPRNN. EURASIP Journal on Audio, Speech, and Music Processing, 2021(1): 1-11. https://doi.org/10.1186/s13636-021-00221-8
[19] Ramdane, M.A., Benallal, A., Maamoun, M., Hassani, I. (2022). Partial update simplified fast transversal filter algorithms for acoustic echo cancellation. Traitement du Signal, 39(1): 11-19. https://doi.org/10.18280/ts.390102
[20] Lashkari, K. (2006). A novel Volterra-wiener model for equalization of loudspeaker distortions. In 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, 5: V-V. https://doi.org/10.1109/ICASSP.2006.1661226
[21] Samuyelu, B., Kumar, P.R. (2019). Robust normalized subband adaptive filter for acoustic noise and echo cancellation systems. Advances in Modelling and Analysis B, 62(2-4): 61-71. https://doi.org/10.18280/ama_b.622-405
[22] Comminiello, D., Scarpiniti, M., Azpicueta-Ruiz, L.A., Arenas-García, J., Uncini, A. (2017). Full proportionate functional link adaptive filters for nonlinear acoustic echo cancellation. In 2017 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, pp. 1145-1149. https://doi.org/10.23919/EUSIPCO.2017.8081387
[23] Wang, D., Chen, J. (2018). Supervised speech separation based on deep learning. An overview, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10): 1702-1726. https://doi.org/10.1109/TASLP.2018.2842159
[24] Graves, A., Mohamed, A.R., Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vancouver, BC, Canada, pp. 6645-6649. https://doi.org/10.1109/ICASSP.2013.6638947
[25] Shi, X., Huang, H., Jian, P., Tang, Y.K. (2021). Improving neural machine translation with sentence alignment learning. Neurocomputing, 420: 15-26. https://doi.org/10.1016/j.neucom.2020.05.104
[26] Goldberg, Y. (2016). A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research, 57: 345-420. https://doi.org/10.1613/jair.4992
[27] Zhou, G.B., Wu, J., Zhang, C.L., Zhou, Z.H. (2016). Minimal gated unit for recurrent neural networks. International Journal of Automation and Computing, 13(3): 226-234. https://doi.org/10.1007/s11633-016-1006-2

[28] Selvanambi, R., Natarajan, J., Karuppiah, M., Islam, S.H., Hassan, M.M., Fortino, G. (2020). Lung cancer prediction using higher-order recurrent neural network based on glowworm swarm optimization. Neural Computing and Applications, 32: 4373-4386. https://doi.org/10.1007/s00521-018-3824-3

[29] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[30] Li, G., Zhang, M., Li, J., Lv, F., Tong, G. (2021). Efficient densely connected convolutional neural networks, Pattern Recognition, 109: 107610. https://doi.org/10.1016/j.patcog.2020.107610

[31] Marti, K. (2008). Stochastic Optimization Methods. Berlin: Springer.

[32] Enzner, G., Buchner, H., Favrot, A., Kuech, F. (2014). Acoustic echo control. In Academic Press Library in Signal Processing, 4: 807-877. https://doi.org/10.1016/B978-0-12-396501-1.00030-3

[33] Burra, S., Kar, A. (2022). Nonlinear stereophonic acoustic echo cancellation using sub-filter-based adaptive algorithm. Digital Signal Processing, 121: 103323. https://doi.org/10.1016/j.dsp.2021.103323

[34] Sankar, S., Kar, A., Burra, S., Swamy, M.N.S., Mladenovic, V. (2020). Nonlinear acoustic echo cancellation with kernelized adaptive filters. Applied Acoustics, 166: 107329. https://doi.org/10.1016/j.apacoust.2020.107329

[35] Comminiello, D., Scarpiniti, M., Azpicueta-Ruiz, L.A., Arenas-García, J., Uncini, A. (2014). Nonlinear acoustic echo cancellation based on sparse functional link représentations. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(7): 1172-1183. https://doi.org/10.1109/TASLP.2014 .2324175

NOMENCLATURE

Acronym	Description
AEC	Acoustic echo cancellation
NAEC	Nonlinear Acoustic Echo Cancellation
SSS	Speech source separation
MM	Magnitude mask
NSM	Near-end signal mask
NL	Nonlinear
ERLE	Echo return loss enhancement
ANN	Artificial Neural Network
CNN	Conventional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short term memory
BILSTM	Bi-Long Short term memory
D^2L^2	Dense LSTM Deep Learning
ASR	Automated speech recognition
LMS	Least mean square
FDKF	Frequency-domain adaptive Kalman filter
STFT	Short-time Fourier transform
RIR	Room impulse response
DFT	Discrete Fourier transform
GRU	Gated recurrent unit
NSAEC	Nonlinear Stereophonic Acoustic echo cancellation
KIPNLMS	Kernel improved proportionate Normalized LMS
PFLAF	Proportionate Functional Link Adaptive Filters