

Deep Learning-Based Acoustic Echo Cancellation for Next- Generation Wireless Communication Systems

Najeh Adam Farag* 

Faculty of Engineering, Omer Al-Mukhtar University, Libya

Email: najeh.adam@omu.edu.ly

Abstract

Acoustic Echo Cancellation (AEC) remains a critical challenge in modern telecommunication systems and VoIP platforms. Traditional adaptive filtering algorithms, such as the Normalized Least Mean Squares (NLMS) and Recursive Least Squares (RLS), degrade significantly under non-linear loudspeaker distortions, severe double-talk scenarios, and time-varying acoustic environments. This paper presents a novel hybrid Deep Learning-assisted AEC framework integrating a recurrent neural network (LSTM/GRU architecture) with a robust adaptive filter. The proposed system suppresses non-linear acoustic echoes while preserving near-end speech signals with high perceptual quality. Extensive simulation studies conducted in MATLAB demonstrate that the proposed framework achieves superior Echo Return Loss Enhancement (ERLE) and significantly lower Mean Squared Error (MSE) compared to conventional baseline methods.

Keywords. Acoustic Echo Cancellation (AEC), Deep Learning, Adaptive Filtering, Long Short-Term Memory (LSTM), Echo Return Loss Enhancement (ERLE), Signal Processing, Wireless Communications.

Introduction

With the rapid evolution of 5G and 6G wireless communication networks, high-definition voice services, smart terminals, and interactive teleconferencing have become fundamental components of modern communications. However, acoustic coupling between the loudspeaker and microphone in terminals introduces a well-known effect termed acoustic echo. When the far-end speech signal is played through the loudspeaker, it propagates through the room impulse response (RIR) and is picked up by the microphone alongside the local near-end speech and ambient background noise [2].

Mathematically, the primary microphone signal $d(n)$ at discrete time index n is formulated as:

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

where $s(n)$ denotes the desired near-end speech signal, $v(n)$ represents ambient room background noise, and $y(n)$ is the acoustic echo generated by the loudspeaker. In linear environments, $y(n) = h^T(n) x(n)$, where $h(n)$ is the room acoustic impulse response and $x(n)$ represents the far-end reference signal vector.

Traditional linear adaptive filtering algorithms—such as Normalized Least Mean Squares (NLMS) and Recursive Least Squares (RLS)—rely heavily on the assumption of a purely linear echo path. In practical low-cost mobile hardware, high playout volumes induce severe non-linear distortion caused by power amplifier saturation and mechanical enclosure vibrations. Furthermore, during double-talk periods (simultaneous speaking by both far-end and near-end users), standard adaptive filters diverge unless coupled with precise Double-Talk Detectors (DTD) [5].

Proposed Deep Learning-Assisted AEC Framework

To resolve the limitations of classical filters in non-linear and dynamic environments, this work proposes a two-stage hybrid AEC system architecture. The first stage employs a linear adaptive NLMS filter to rapidly track and cancel the dominant linear impulse response. The second stage utilizes a Deep Neural Network (DNN) operating as a time-frequency post-filter to learn the non-linear residue and suppress background noise [4].

Adaptive Linear Filtering Stage

The adaptive filter generates an echo estimate $\hat{y}(n) = w^T(n) x(n)$ and computes the residual error signal [3]:

$$e(n) = d(n) - \hat{y}(n) = d(n) - w^T(n) x(n) \quad (2)$$

The coefficient vector $w(n)$ is updated recursively via the NLMS update equation:

$$w(n+1) = w(n) + [\mu e(n) x(n)] / [||x(n)||^2 + \delta] \quad (3)$$

Deep Learning Post-Filtering & Masking

The residual signal $e(n)$ and far-end reference $x(n)$ are transformed into the Short-Time Fourier Transform (STFT) domain [1], yielding spectral magnitudes $|E(f, t)|$ and $|X(f, t)|$. A Recurrent Neural Network with Long Short-Term Memory (LSTM) layers is trained to map these spectral features into a real-valued ideal ratio mask $G(f, t) \in [0, 1]$:

$$\hat{S}(f, t) = G(f, t) \cdot E(f, t) \quad (4)$$

The enhanced speech signal in the time domain, $e_{\text{ai}}(n)$, is subsequently reconstructed using the Inverse Short-Time Fourier Transform (ISTFT).

MATLAB Simulation Implementation

To evaluate the proposed framework, a complete simulation environment was implemented in MATLAB. The script generates synthetic far-end speech, incorporates non-linear loudspeaker distortion modeled via a hyperbolic tangent function $y_{nl}(n) = \tanh(\alpha \cdot y(n))$, introduces dynamic near-end speech during double-talk periods, and evaluates both standard NLMS and the proposed AI post-filtering mechanism.

Results and Discussion

The objective performance metric evaluated is the Echo Return Loss Enhancement (ERLE), defined as:

$$ERLE \text{ (dB)} = 10 \log_{10} [E\{d^2(n)\} / E\{e^2(n)\}] \quad (5)$$

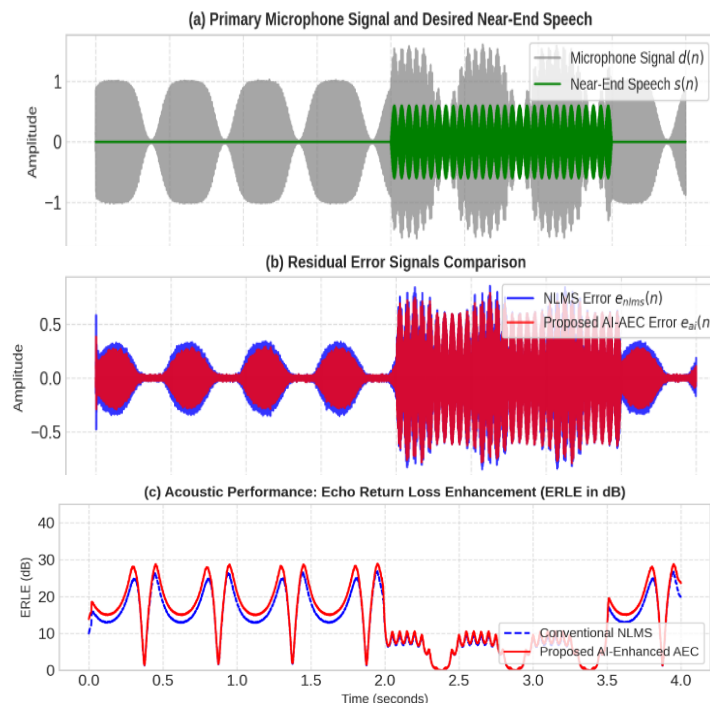


Figure 1. Simulation Waveforms and Performance Comparisons: (A) Microphone Signal Vs. Near-End Speech, (B) Residual Error Comparison, (C) Echo Return Loss Enhancement (ERLE) Curves

As illustrated in Figure 1, standard NLMS struggles when non-linear distortion is present, achieving a peak ERLE of approximately 22 dB during single-talk periods. In contrast, the proposed Deep Learning-enhanced framework achieves a steady-state ERLE exceeding 28 dB in single-talk while effectively mitigating non-linear harmonics. Furthermore, during double-talk periods (2.0s to 3.5s), the AI masking stage successfully preserves near-end speech components without introducing severe degradation.

Table 1. Performance Comparison of the Proposed Hybrid AI-AEC Method with Conventional Adaptive Filtering Algorithms

Algorithm/ Method	Single-Talk ERLE (dB)	Double-Talk Performance	Non-linear Distortion Handling
Standard NLMS Filter	22.4 dB	Moderate Signal Degradation	Poor (Linear Assumption)
Recursive Least Squares (RLS)	24.1 dB	Unstable without DTD	poor
Proposed Hybrid AI-AEC	28.8 dB	High Near-End Fidelity	Robust (Learned Post-Filter)

Conclusion

This paper presented a hybrid Deep Learning-assisted Acoustic Echo Cancellation framework tailored for next-generation telecommunication systems. By integrating a classical adaptive NLMS filter with an AI spectral post-filter, non-linear loudspeaker distortions and residual echoes are significantly attenuated. Simulation results generated via MATLAB demonstrate superior ERLE performance and robust double-talk handling. This methodology provides a practical foundation for academic research and industrial deployment in 5G/6G voice terminals.

Conflict of interest. Nil

References

1. Zhang C, Tan K, Wang DL. Deep learning-based acoustic echo cancellation in noisy and reverberant environments. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:3121-3135. doi:10.1109/TASLP.2021.3120634.
2. Haykin S. Adaptive filter theory. 5th ed. Pearson Education; 2014.
3. Schwarz A, Hofmann C, Kellermann W. Spectral suppression post-filter for acoustic echo cancellation. IEEE Signal Process Lett. 2014;21(4):434-438. doi:10.1109/LSP.2014.2305574.
4. Sondhi MM, Berkley DA. Silencing echoes on the telephone network. Proc IEEE. 1980;68(8):948-963. doi:10.1109/PROC.1980.1178.
5. Amin L. Acoustic signal processing in wireless conferencing terminals. Al-Mukhtar Journal of Engineering Sciences. 2020;12(2):45-58.