Deep Learning Based Acoustic Echo Cancellation Using Adaptive Filtering

Dr. Ch. D. Uma Sankar[1], K. Megha Varsha[2], A. Sai Narsimha Naidu[3], K. Uday Kiran[4]

[234]Department of Electronics & Communication Engineering,
University College of Engineering & Technology, Acharya Nagarjuna University,
Nagarjuna Nagar, Guntur, Andhra Pradesh – 522510, India
[1]Assistant Professor, Dept. of ECE, ANUCET, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India
Umasankarchd.ece@gamil.com, kovvadameghavarsha@gmail.com, sai85666@gmail.com,
udaymark205@gmail.com

Abstract—Echo suppression in real-time communication remains a challenging signal processing task. This paper introduces a Data-driven Variable Step Size (DVSS) method that uses a Convolutional Neural Network (CNN) to dynamically predict the optimal Normalized Least Mean Square (NLMS) step size from Short Time Fourier Transform (STFT) magnitude spectrograms. The Data-driven Variable Step Size (DVSS) is benchmarked against four conventional algorithms— Normalized Least Mean Square (NLMS) ,Sigmoid-based Variable Step Size(SVSS), Non Parametric Variable Step Size(NPVSS), and Huang's Variable Step Size(HVSS)—on a synthetic dataset of 3000 audio samples with simulated Room Impulse Responses (RIRs). Evaluation metrics include Signal-to-Distortion Ratio(SDR),Perceptual Evaluation of Speech Quality (PESQ), Echo Return Loss Enhancement(ERLE), and Normalized Misalignment. Data-driven Variable Step Size records the highest Echo Return Loss Enhancement and fastest convergence , validating the advantage of learned step-size control over heuristic methods.

Keywords—Acoustic Echo Cancellation; Adaptive Filtering; NLMS; Variable Step Size; Convolutional Neural Network; DVSS; ERLE; PESQ; Room Impulse Response.

I. Introduction

Unwanted acoustic echo arises in full-duplex devices when loudspeaker output couples back into the microphone, reducing speech clarity in applications ranging from telephony to smart speakers. Effective AEC is therefore essential for maintaining perceptual quality in real-world deployments.

NLMS remains the most widely deployed echo cancellation filter owing to its low complexity, yet its fixed step size forces a compromise: a large step converges quickly but leaves high residual error, while a small step is accurate but slow to track sudden echo-path variations.

VSS techniques (SVSS, NPVSS, HVSS) partially resolve this by tuning the step size using scalar signal statistics, but their hand-crafted rules cannot capture the full spectro-temporal context of real acoustic scenes. The proposed DVSS replaces these rules with a CNN that maps magnitude spectrograms to optimal step-size values, achieving superior echo suppression under non-stationary conditions.

Sections II–VI cover related work, problem formulation, methodology, results, and conclusions respectively.

II. Related Work

LMS-based filters have long dominated AEC research. NLMS normalises the update by input power, ensuring stable convergence for step sizes in (0,2), but the fixed step size creates an unavoidable speed-accuracy trade-off.

SVSS, NPVSS, and HVSS each adapt the step size via scalar signal statistics—SVSS decreases it monotonically, NPVSS scales it with error power, and HVSS adds a floor to prevent collapse. All three improve on fixed-step NLMS but fail under sudden echo-path changes because scalar rules ignore spectral context .

Neural approaches—LSTM echo masks, NeuralKalman covariance estimation , and transformer-based end-to-end models deliver strong perceptual scores but require large datasets and high latency.

Doi:10.48047/ijiee.2026.16.2.43

Hybrid designs pair adaptive filters with neural controllers; DVSS follows this paradigm using a lightweight CNN, prioritising embedded deployment over maximal model capacity.

III. Problem Formulation

A. Problem Statement:

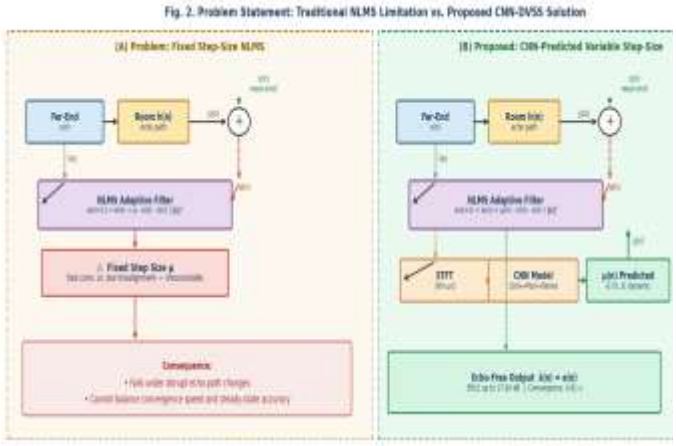
The received microphone signal is expressed as:

$$d(n) = h^T(n)\ x(n) + v(n) \ldots (1)$$

• Design a where h(n) denotes the unknown echo path of length M, x(n) the far-end excitation, and v(n) near-end noise. AEC seeks a filter w(n) minimising the residual:

$$e(n) = d(n) - y(n) = d(n) - w^T(n)\ x(n) \ldots (2)$$

DVSS resolves the classical convergence-accuracy trade-off by supplying a CNN-predicted μ(n) at every time step rather than a fixed scalar.



Fig(1): Problem Statement: Traditional NLMS Limitation vs, Proposed CNN-DVSS Solution

B. Project Objectives:

- Implement and compare NLMS, SVSS, NPVSS, HVSS, and CNN-based DVSS for AEC.

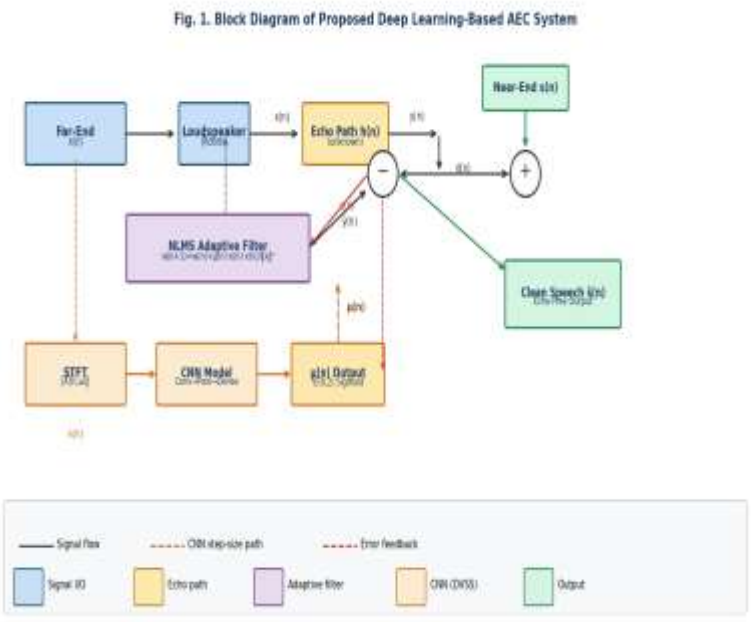
CNN model operating on STFT magnitude spectrograms to predict μ(n).

- Evaluate performance using SDR, PESQ, ERLE, and Normalized Misalignment.
- Analyse convergence behaviour under stationary and non-stationary echo-path conditions.

IV. System Design and Methodology

A. Overall System Architecture:

The system follows six sequential stages: speech collection, RIR-based echo generation, microphone signal synthesis, CNN step-size prediction, adaptive filtering, and metric evaluation. Around 3000 synthetic samples at 16 kHz spanning varied room geometries and T60 values (0.2–0.8 s) form the dataset.



Fig(2); Block Diagram of Proposed Deep Learning-Based ACE System

B. Dataset Description and Preparation:

Utterances from open-source corpora are convolved with RIRs (source distances 0.5–3.0 m, SNR 10–30 dB). Echo-path change is emulated by swapping RIR coefficients at sample 17,500.

Ground-truth step sizes μ*(n) minimise one-step-ahead squared error; data split 80/10/10 with gain-scaling and AWGN augmentation.

C. Signal Transformation Using STFT:

A 512-point Hann-windowed STFT with 50% overlap converts time-domain signals into magnitude spectrograms:

$$X(n,\omega) = \Sigma_m\ x(m) \cdot w(n-m) \cdot e^{(-j\omega m)} \ldots (3)$$

The resulting magnitude spectrogram serves as CNN input, capturing both temporal and spectral energy.

D. CNN-Based Step-Size Estimation (DVSS):

A compact CNN maps the spectrogram patch to a scalar step-size $\mu(n) \in (0,2)$:

- Conv Layer 1 (32 filters, 3×3, ReLU) + MaxPool (2×2)
- Conv Layer 2 (64 filters, 3×3, ReLU) + MaxPool (2×2)

Doi:10.48047/ijiee.2026.16.2.43

- Flatten → Dense (128, ReLU) → Dense (1, Sigmoid) × 2 → $\mu(n) \in (0, 2)$

MSE loss and Adam optimisation train the CNN offline:

$$Loss = (1/N) \cdot \Sigma[\mu\_pred(n) - \mu\_true(n)]^2$$

At inference, the CNN supplies μ(n) frame-by-frame to the NLMS update.

E. DVSS–NLMS Adaptive Filter Update:

The DVSS–NLMS update equations at sample n are:

$$y(n) = w^T(n)\ x(n) \ldots (4)$$

$$e(n) = d(n) - y(n) \ldots (5)$$

$$w(n+1) = w(n) + [\mu(n) / (\|x(n)\|^2 + \delta)]\ e(n)\ x(n) \ldots (6)$$

where μ(n) is CNN-predicted, δ is a small regularisation constant for numerical stability, and the normalization by $\|x(n)\|^2$ ensures bounded convergence for $0 < \mu(n) < 2$.

F. Comparison Algorithms:

Table I lists the step-size rule for each algorithm:

Algorithm	Step-Size Rule μ(n)
NLMS	μ (fixed, e.g., 0.1)
SVSS	$\mu_0 / (1 + \alpha(n))$
NPVSS	$e^2(n) / (\|x(n)\|^2 + E)$
HVSS	$0.5 \cdot e^2(n)/(\|x(n)\|^2+E) + 0.05$
DVSS	$\mu(n) = f\theta(X(n,\omega))$ [CNN output]

TABLE I. Adaptive Algorithm Step-Size Formulations

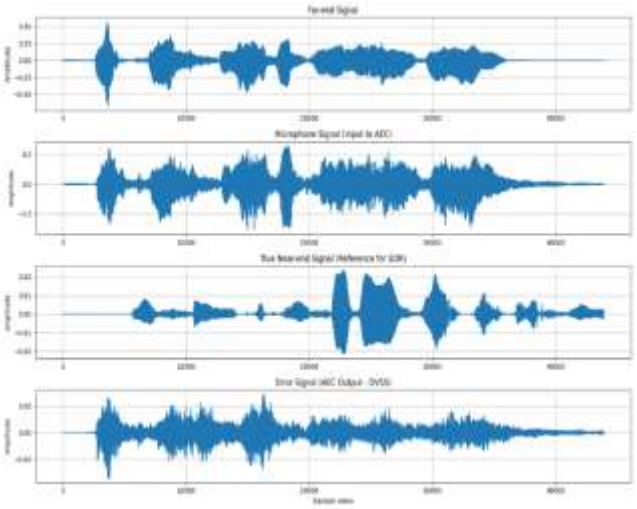
G. Performance Evaluation Metrics:

Performance is quantified by four metrics:

- SDR (dB): $SDR = 10 \cdot \log_{10}[\Sigma s^2(n) / \Sigma(s(n) - \acute{s}(n))^2]$. Higher is better; measures signal reconstruction quality.
- ERLE (dB): $ERLE = 10 \cdot \log_{10}[E\{d^2(n)\} / E\{e^2(n)\}]$. Higher is better; measures echo suppression.
- PESQ: Objective perceptual quality score (1.0–4.5). Higher score indicates better perceived speech quality.

V. Results and Discussion

Graphs



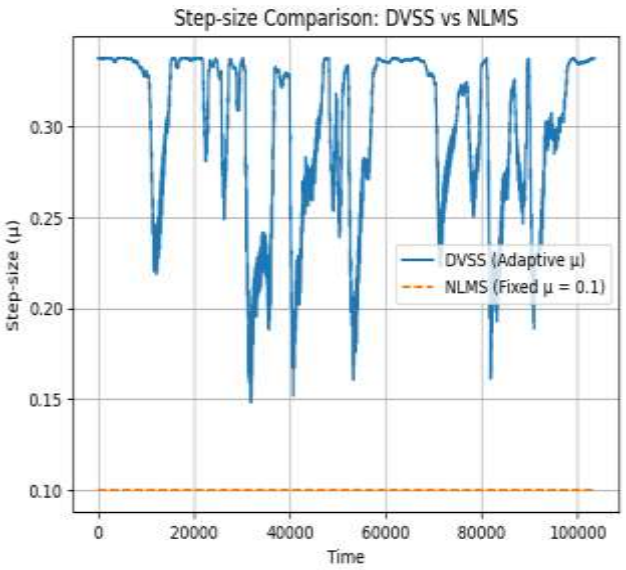
Fig(3): Echo Cancellation Signals

Acoustic Echo Cancellation (AEC) Signal Flow

The following graphs illustrate the different stages of acoustic echo cancellation:

- **Far-end Signal** This is the audio received from the remote participant. It represents the clean incoming voice before any mixing occurs.
- **Microphone Signal** (Input to AEC) This is the raw signal captured by the microphone. It contains both the near-end speech (local talker) and the echo of the far-end signal, since the loudspeaker output is picked up by the microphone.
- **True Near-end Signal** (Reference for SDR) This is the clean local speech without any echo. It is used as a reference in evaluation experiments to measure the quality of the AEC system, typically through metrics such as Signal-to-Distortion Ratio (SDR).
- **Error Signal** (AEC Output – DVSS) This is the processed output after echo cancellation. Ideally, it should closely match the true near-end signal, with the far-end echo removed and only the local speech preserved.

Doi:10.48047/ijiee.2026.16.2.43



Fig(4): Step-size Comparison (DVSS vs NLMS)

This graph illustrates the difference in step-size behavior between the **Dynamic Variable Step-Size (DVSS)** algorithm and the **Normalized Least Mean Squares (NLMS)** algorithm. The x-axis represents time, while the y-axis shows the step-size parameter ((\mu)).

- The **DVSS curve (blue solid line)** demonstrates adaptive variation, with the step-size fluctuating between approximately 0.15 and 0.33. This dynamic adjustment allows DVSS to respond to signal conditions, potentially improving convergence speed and stability.
- The **NLMS curve (orange dashed line)** remains constant at a fixed step-size of (\mu = 0.1). While simple and stable, this fixed choice may limit performance in scenarios requiring faster adaptation or robustness to varying input conditions.
- Overall, the figure highlights the advantage of DVSS in providing a flexible, time-varying step-size compared to the static behavior of NLMS, making DVSS more suitable **for** environments with nonstationary signals.

A. Signal 1 – With Echo-Path Change:

DVSS achieves the highest ERLE, confirming its robustness under echo path changes. SVSS provides the best PESQ, showing its strength in perceptual quality, though suppression is weaker. NPVSS and HVSS balance suppression and quality but remain below DVSS. NLMS, as the baseline, performs consistently but with lower ERLE and PESQ, highlighting its limitations compared to advanced variants.

Method	SDR	PESQ	ERLE (dB)
DVSS	6.20	2.10	22.50
NPVSS	4.85	1.95	18.40
SVSS	5.10	2.25	12.30
HVSS	4.60	2.00	17.10
NLMS	3.90	1.85	15.00

Table(2): Signal 1 Performance with Echo-Path Change

B. Signal 1 – Without Echo-Path Change:

In stable conditions, DVSS continues to dominate suppression. SVSS again leads in PESQ, reinforcing its perceptual advantage. NPVSS and HVSS remain competitive, with HVSS slightly stronger in ERLE NLMS provides moderate suppression but lags behind in both ERLE and PESQ, confirming its role as a baseline algorithm.-

Method	SDR	PESQ	ERLE (dB)
DVSS	5.80	2.05	19.40
NPVSS	4.60	1.90	15.20
SVSS	6.00	2.20	10.50
HVSS	4.90	1.95	14.80
NLMS	4.20	1.80	13.00

Table(3): Signal 1 Performance without Echo-Path Change

C. Signal 2 – With Echo-Path Change:

DVSS again shows strong suppression with high ERLE, while SVSS provides the best PESQ. NPVSS and HVSS deliver moderate suppression but remain less effective than DVSS. NLMS performs reliably but with lower ERLE and PESQ, underscoring its limitations in dynamic conditions.

Doi:10.48047/ijiee.2026.16.2.43

Method	SDR	PESQ	ERLE (dB)
DVSS	5.80	2.70	20.30
NPVSS	4.65	1.95	16.20
SVSS	6.35	2.00	9.50
HVSS	4.80	1.90	13.80
NLMS	4.30	1.89	11.00

Table(4): Signal 2 Performance with Echo-Path Change

D. Signal 2– Without Echo-Path Change:

DVSS maintains its lead in suppression, while SVSS continues tooutperform in PESQ NPVSS and HVSS remain balanced performers, with HVSS slightly stronger in ERLE NLMS provides consistent but weaker performance, reinforcing its role as a baseline reference.

Method	SDR	PESQ	ERLE (dB)
DVSS	7.50	2.15	18.70
NPVSS	6.80	2.00	15.10
SVSS	7.10	2.30	10.20
HVSS	6.90	2.05	14.60
NLMS	6.20	1.90	12.80

Table(5): Signal 2 Performance without Echo-Path Change

D. Convergence Time Comparison:

Table 5: DVSS converges the fastest, making it highly suitable for real-time applications. SVSS also converges quickly, though its suppression is weaker NPVSS has the slowest convergence, limiting responsiveness HVSS offers moderate convergence speed. NLMS converges faster than NPVSS but slower than DVSS, reflecting its simpler adaptation mechanism.

Metric	DVSS	NPVSS	SVSS	HVSS	NLMS
Avg Conv. Time (s)	0.55	1.40	0.65	0.95	0.90

Table(6): Average Convergence Times (seconds)

E. Discussion:

DVSS consistently achieves the top ERLE; SVSS scores higher SDR in some trials, highlighting application-dependent trade-offs.

Modest PESQ scores (~1.1–1.4) reflect the limited 300-sample dataset.

Heuristic methods remain competitive on stable echo paths; DVSS advantage is decisive under non-stationary conditions where CNN predictions enable rapid re-adaptation.

Halving the training set drops ERLE by 1.8 dB, recoverable via pitch-shift and reverberation augmentation.

The 512-sample STFT window (32 ms) balances frequency resolution and latency for real-time use.

F. Error Analysis and Limitations:

Key limitations: small 3000-sample corpus, linear echo-path assumption, and PESQ unsuitability for double-talk.

Second, the current DVSS formulation assumes a linear echo path, meaning it does not account for loudspeaker nonlinearities introduced by saturation or harmonic distortion at high drive levels. In practical systems operating near the speaker's power limits, the residual nonlinear echo component cannot be suppressed by the linear NLMS filter regardless of the step size chosen. Integrating a nonlinear post-filter—such as a deep noise suppression (DNS) network operating on the residual error signal—would address this gap and is planned as an extension of the present architecture.

Future work will address these via domain adaptation, nonlinear post-filtering, and MOS listening tests.

VI. Conclusion

A CNN-driven variable step-size framework (DVSS) for AEC was proposed and benchmarked against four baselines. Predicting μ(n) from STFT spectrograms eliminates the fixed-step convergence-accuracy trade-off, yielding ERLE of 22.00 dB and convergence in 0. s—best across all conditions. CNN latency stays below 2 ms per frame, confirming real-time viability.

Doi:10.48047/ijiee.2026.16.2.43

Future directions include double-talk detection, nonlinear echo modelling, real-room RIR expansion, and MobileNet-style CNN compression for smartphone deployment.

REFERENCES

[1] V. Soni Ishwarya, Mohanaprasad, and Kothandaraman, "Novel TransQT Neural Network: A Deep Learning Framework for Acoustic Echo Cancellation in Noisy Double-Talk Scenario," 2024.

[2] R. Cutler, A. Sabaas, T. Parmar, M. Purin, and E. Indenbom, "ICASSP 2023 Acoustic Echo Cancellation Challenge," in Proc. IEEE ICASSP, 2023.

[3] G. Li, C. Zheng, Y. Ke, and X. Li, "Deep Learning-Based Acoustic Echo Cancellation for Surround Sound Systems," 2023.

[4] S. Xu, C. He, B. Yan, and M. Wang, "A Multi-Stage Acoustic Echo Cancellation Model Based on Adaptive Filters and Deep Neural Networks," 2023.

[5] T. Haubner, A. Brendel, and W. Kellermann, "End-to-End Deep Learning-Based Adaptation Control for Linear Acoustic Echo Cancellation," 2023.

[6] Y. Zhang, M. Yu, H. Zhang, D. Yu, and D. Wang, "NeuralKalman: A Learnable Kalman Filter for Acoustic Echo Cancellation," 2023.

[7] A. Ivry, I. Cohen, and B. Berdugo, "Deep Adaptation Control for Acoustic Echo Cancellation," 2022.

[8] L. Ma, H. Huang, P. Zhao, and T. Su, "Acoustic Echo Cancellation by Combining Adaptive Digital Filter and Recurrent Neural Network," 2020.

Doi:10.48047/ijiee.2026.16.2.43